# Fungi use highly diverse approaches for plant biomass conversion as revealed through bioinformatic analysis

Jiajia Li

# Fungi use highly diverse approaches for plant biomass conversion as revealed through bioinformatic analysis

**Jiajia Li**

**September 2023**

Jiajia Li

Fungi use highly diverse approaches for plant biomass conversion as revealed through bioinformatic analysis

PhD thesis, Utrecht University, Utrecht, the Netherlands (2023)

ISBN: 978-94-6469-551-9

Cover design: Jiajia Li

Lay-out design: Jiajia Li

Printing: ProefschriftMaken || www.proefschriftmaken.nl

# Fungi use highly diverse approaches for plant biomass conversion as revealed through bioinformatic analysis

**Schimmels gebruiken zeer diverse strategieën voor de conversie van plantenbiomassa, ontrafelt door bioinformatische analyse**
(met een samenvatting in het Nederlands)

**Proefschrift**

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 26 september 2023 des middags te 16.15 uur

door

**Jiajia Li**

geboren op 9 januari 1992
te Shanxi, China

**Promotoren**:
Prof. dr. R.P. de Vries
Prof. dr. B. Snel

**Copromotor**:
Dr. M. Peng

**Beoordelingscommissie**:
Prof. dr. A.F.J.M. Van Den Ackerveken
Prof. dr. B. Teusink
Prof. dr. A. Tsang
Prof. dr. ir. M.W. Fraaije
Dr. I. Druzhinina

*To my family, who are always there to support and accompany me through this journey*

The research described in this thesis was performed at the Fungal Physiology Group, Westerdijk Fungal Biodiversity Institute (formerly CBS-KNAW), Utrecht, the Netherlands and supported by the China Scholarship Council (grant no. 201909110079).




# CONTENTS

# Chapter 1

# General introduction

Chapter 1

# 1. Plant biomass composition and its conversion

Plant biomass is a magnificent renewable resource and therefore is of major importance for ecology and the global carbon cycle. In nature, fungi play central roles in the degradation of plant biomass, as they are highly efficient degraders of plant polysaccharides. Plant biomass degradation by filamentous fungi has been the subject of numerous studies for decades. The use of fungi and their encoded enzymes to produce high value products from plant biomass holds great potential for transforming current fossil-based economies into a sustainable bioeconomy [1]. For instance, fungal biotechnology has been widely used in many industrial applications, such as baking, beverages, animal feed, pulp and paper, textiles, detergents, production of bio-fuels and bio-chemicals [2, 3].

Plant cell walls are the major fraction of the plant biomass and mainly consist of plant polysaccharides, proteins and lignin [4]. Plant polysaccharides can also be classified as plant cell wall polysaccharides (e.g., cellulose, hemicellulose and pectin) and storage polysaccharides (e.g., starch, gums, inulin) [5]. The chemical composition and molecular structure of plant biomass vary dramatically depending on their origin, organ type and harvest season.

To degrade plant biomass, fungi have evolved a fine-tuned system for the efficient conversion of specific types of plant biomass. This process involves several key cellular activities, including nutrient sensing, secretion of degrading enzymes (Carbohydrate-Active enZymes, CAZy), sugar uptake by transporters, intracellular metabolism, and the sophisticated transcription regulation system for precisely governing all aspects of the process (Figure 1). As depicted in Figure 1, a monomeric or short oligomeric component of a polysaccharide is transported into the cell, resulting in the activation of specific transcription factors (regulators) that then enter the nucleus and bind to the promoter of their target genes. These regulators promote the expression of their target genes to activate specific sets of sugar metabolic and extracellular plant biomass degrading enzymes, resulting in liberation and conversion of

more of the monomeric compounds from plant polymers. Therefore, a better understanding of the molecular mechanisms and key genes underlying fungal plant biomass conversion and investigating their evolutionary diversity across different species is crucial for further exploring the ecological roles of fungi and developing plant biomass related fungal biotechnology.

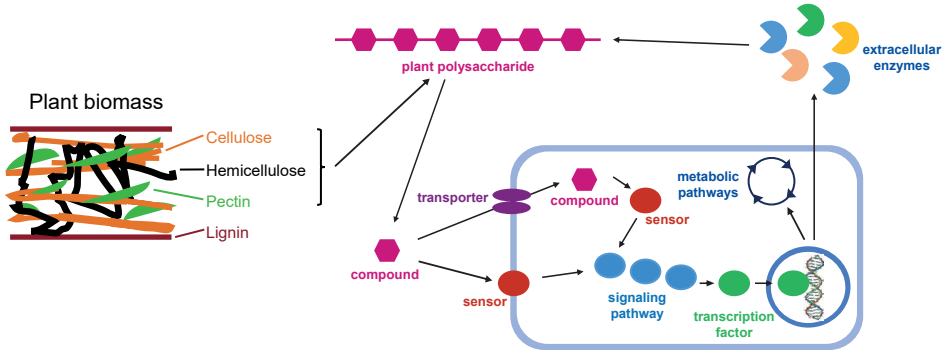


**Figure 1. Conversion of plant polysaccharides by fungi**. The figure is modified from [5].

## 2. Plant polysaccharides-degrading enzymes

Given the complex structure of plant polysaccharides, a broad range of extracellular enzymes is required to break the different linkages of plant polysaccharides. In the well-known CAZy (Carbohydrate-Active Enzyme) database (http://www.cazy.org/) [6, 7], most of these plant degrading enzymes have been classified into specific enzyme families according to the catalytic domains and carbohydrate binding modules present in their amino acid sequences. Currently, the CAZy database has defined five major enzyme classes: glycoside hydrolases (GHs), carbohydrate esterases (CEs), polysaccharide lyases (PLs), glycosyltransferases (GTs), and auxiliary activities (AA). Numerous families can be identified within each of these classes, but only a subset of enzymes in the CAZy database have been experimentally shown to be plant biomass degrading enzymes. Some families contain only a single (known) enzyme activity (e.g. GH67 [8] and CE8, while other families contain multiple activities (e.g. GH43 [9, 10], GH28 [11, 12],

and GH30 [13]). A continuous effort is required to further explore the less well-studied CAZymes families, as well as the functional diversity within specific families [14]. Below we briefly summarize the CAZy families that are involved in plant polysaccharide degradation (Table 1 and Figure 2).

### *2.1 Cellulose degradation*

Cellulose is the most abundant plant cell wall polysaccharide, with linear chains of β-1,4-linked D-glucose. Four enzyme groups are involved in the biodegradation of cellulose: endoglucanases (EGLs), cellobiohydrolases (CBHs), β-glucosidases (BGLs), and auxiliary activities (AAs) [15]. EGLs and CBHs hydrolyze cellulose into gluco-oligosaccharides and cellobiose [16, 17], respectively. Subsequently, cellobiose is cleaved into D-glucose units by BGLs. Cellulose-active lytic polysaccharide monooxygenases (LPMOs) from AA9 and AA16 can boost cellulose degradation by increasing the substrate availability and accessibility for EGLs and CBHs [18]. In addition, this LPMO-boosting effect is suggested to benefit from synergy with cellobiose dehydrogenases (CDHs, AA3_1) [19-23]. The enzymes involved in the complete biodegradation of cellulose are indicated in Table 1 and Figure 2.

### *2.2 Hemicellulose degradation*

Hemicellulose is the second most abundant polymeric constituent in plant biomass, which is strongly associated with cellulose and plays an important role in strengthening plant cell walls [24]. It is present in three structurally different polysaccharide forms: xylans, xyloglucans and mannans. Xylans have a β-1,4-linked D-xylose backbone and may also contain other monosaccharide substituents like D-galactose, D-xylose, L-arabinose, and D-glucuronic acid. Xyloglucan has the same backbone as cellulose but is decorated with α-1,6-linked D-xylose residues, and additionally contains D-galactose, L-fucose, or L-arabinose residues [25]. Mannans consist of a β-1,4-linked D-mannose backbone with D-galactose side chains and its backbone can be interrupted by varying amounts of β-1,4-linked D-glucose residues. Hydrolysis and

degradation of hemicelluloses requires a multi-enzyme system due to the different types of backbones and side groups. A variety of hemicellulolytic enzymes have been identified, including at least 20 GH families, four CE families and FAE (Table 1 and Figure 2).

Xylan degradation

Xylans are sometimes referred to as arabinoxylan, glucuronoxylan or glucuronoarabinoxylan, depending on the substitution of the xylose backbone. Due to its variable composition, xylan requires a larger number of enzymatic activities to achieve full degradation than cellulose (Table 1). Endo-β-xylanase (XLN) and β-D xylobiohydrolase (XBH) are glycoside hydrolases (GHs) able to cleave the xylan backbone [26, 27], while β-D-xylosidase (BXL) acts on oligosaccharides. Depending on the types of xylan, various xylan side chain degrading enzymes are required for the full xylan degradation, mainly including α-arabinofuranosidase (ABF), α-glucuronidase (AGU), glucuronoyl esterase (GE), acetyl xylan esterase (AXE), hemicellulose acetyl esterase (HAE) [28], arabinoxylan arabinofuranohydrolases (AXH), and feruloyl esterase (FAE) [27] (Table 1). They belong to different classes, i.e., GH and carbohydrate esterase (CE).

Xyloglucan degradation

Xyloglucan, contains a β-1,4-glucan backbone that is regularly decorated with other sugar residues. Based on different side chain structures [29], the complete hydrolysis of xyloglucan requires a vast arsenal of enzymes with various activities and specificities (Table 1). For example, the linkages between D-xylose and D-glucose are hydrolyzed by α-xylosidases (AXL), those between D-galactose and D-xylose by β-1,4-galactosidases (LAC) and those between L-fucose and D-galactose by α-fucosidases (AFC). L-arabinose residues can be removed by α-L-arabinofuranosidases (ABF). The degradation of xyloglucan involves nine GH families: GH29, GH95 [11], GH141, GH12 [30], GH44 [31], GH74 [32], GH31 [11], GH2 and GH35 [15].

**Table 1. CAZy families involved in plant polysaccharide degradation.**

| Polysaccharide | | Enzyme activity | Abbreviation | CAZy families | Monomers |
|---|---|---|---|---|---|
| **Cellulose** | | β-1,4-endoglucanase | EGL | **GH** 5_4, 5_5, 5_22, 12, 45, 131; | D-glucose |
| | | cellobiohydrolases | CBH | **GH** 6, 7; | |
| | | β-glucosidases | BGL | **GH** 1, 3; | |
| | | lytic polysaccharide monooxygenases | LPMO | **AA** 9, 16; | |
| | | cellobiose dehydrogenase | CDH | **AA** 3_1; | |
| **Hemicellulose** | **Xylan(s)** (Xylan, Glucuronoxylan, Glucuronoarabinoxylan, Arabinoxylan) | α-arabinofuranosidase | ABF | **GH** 43, 51, 54; | D-xylose, L-arabinose, D-galactose, D-glucuronic acid; |
| | | glucuronidase | AGU | **GH** 67, 115**;** | |
| | | acetyl xylan esterases | AXE | **CE** 1, 5**;** | |
| | | arabinoxylan arabinofuranohydrolase | AXH | **GH** 62; | |
| | | β-1,4-xylosidase | BXL | **GH** 3, 43; | |
| | | glucuronoyl esterase | GE | **CE** 15; | |
| | | Endoxylanase | EXL | **GH** 30_7; | |
| | | endo-β-1,4-xylanase | XLN | **GH** 10, 11; | |
| | | hemicellulose acetyl esterases | HAE | **CE** 16**;** | |
| | | feruloyl esterase | FAE | \ | |
| | **Xyloglucan** | xyloglucan β-1,4-endoglucanase | XEG | **GH** 12, 44, 74; | D-glucose, D-galactose, L-fucose, D-xylose, (L-arabinose); |
| | | α-arabinofuranosidase | ABF | **GH** 43, 51, 54; | |
| | | α-fucosidase | AFC | **GH** 29, 95, 141; | |
| | | α-xylosidase | AXL | **GH** 31; | |
| | | β-1,4-galactosidase | LAC | **GH** 2, 35; | |
| | | hemicellulose acetyl esterases | HAE | **CE** 16; | |
| | **Mannan(s)** | α-1,4-galactosidase | AGL | **GH** 27, 36; | D-mannose, D-galactose, D-glucose, D-glucuronic acid; |
| | | β-1,4-endomannanase | MAN | **GH** 5_7, 26, 134; | |
| | | β-1,4-mannosidase | MND | **GH** 2**;** | |
| | | hemicellulose acetyl esterases | HAE | **CE** 16; | |

(**Table 1.** Continued)

| Polysaccharide | Enzyme activity | Abbreviation | CAZy families | Monomers |
|---|---|---|---|---|
| **Pectin (HG, XGA, RG-I, RG-II)** | endopolygalacturonases/ exopolygalacturonases/endorhamnogalacturonase/ exo-rhamnogalacturonase/ xylogalacturonase | PGA/PGX/ RHG/RGX/ XGH | **GH** 28; | D-galacturonic acid, D-galactose, L-arabinose, L-rhamonose, D-glucuronic acid, D-xylose, D-glucose, L-fucose |
| | α-arabinofuranosidase | ABF | **GH** 43, 51, 54; | |
| | endoarabinanase | ABN | **GH** 43; | |
| | β-1,4-endogalactanase | GAL | **GH** 53; | |
| | α-rhamnosidase | RHA | **GH** 78; | |
| | unsaturated glucuronyl hydrolase | UGH | **GH** 88; | |
| | exoarabinanase | ABX | **GH** 93; | |
| | unsaturated rhamnogalacturonan hydrolase | URGH | **GH** 105; | |
| | endo-β-1,6-galactanase | \ | **GH** 5_16; | |
| | pectin lyase | PEL | **PL** 1; | |
| | pectate lyase | PLY | **PL** 1_2, 1_4, 1_7, 3_2, 9_3; | |
| | rhamnogalacturonan lyase | RGL | **PL** 4_1, 4_3, 11; | |
| | pectin methyl esterase | PME | **CE** 8**;** | |
| | rhamnogalacturonan acetyl esterases | RGAE | **CE** 12**;** | |
| | feruloyl esterase | FAE | \ | |
| | β-1,4-galactosidase | LAC | **GH** 2, 35**;** | |
| | lytic polysaccharide monooxygenases | LPMO | **AA** 17; | |
| **Starch (Amylose, Amylopectin)** | α-glucosidase | AGD | **GH** 13_40, 31; | D-glucose |
| | amylo-α-1,6-glucosidase | AMG | **GH** 133**;** | |
| | α-amylase | AMY | **GH** 13_1, 13_5, 13_8; | |
| | glucoamylase | GLA | **GH** 15**;** | |
| | lytic polysaccharide monooxygenases | LPMO | **AA** 13**;** | |
| **Inulin (β-1,2-fructans)** | exdo-inulinase | INU | **GH**32 | D-glucose, D-fructose |
| | exo-inulinase | INX | | |
| | invertase | INV | | |

Mannan degradation

Mannanases are vital enzymes involved in the hydrolysis of mannan, mostly including β-1,4-endomannanase (MAN), β-1,4-mannosidase (MND), β-1,4-glucosidases, α-galactosidases (AGL) and hemicellulose acetyl esterases (HAE). Fungal AGLs have been assigned to GH27 and GH36. Most of the fungal MANs are classified in GH5_7 and fewer in GH26 and GH134. MND and HAE belong to GH2 [33] and CE16, respectively (Table 1).

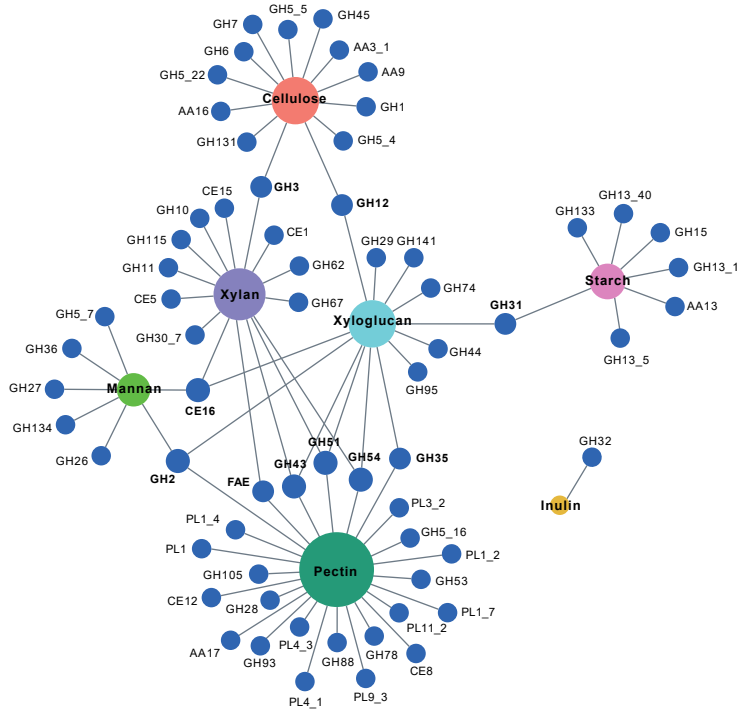


**Figure 2. CAZymes families related to plant biomass degradation**. Families in bold are involved in the degradation of multiple polysaccharides. The figure was copied from **Chapter 4**.

### *2.3 Pectin degradation*

Pectins are the major components of the primary cell wall, and they are present in three main substructures that are rich in D-galacturonic acid, including: homogalacturonan (HG), which is a linear homopolymer of α-1,4-linked

D-galacturonic acid residues; xylogalacturonan (XGA), which is an HG substituted with β-1,3-linked D-xylose; and two types of rhamnogalacturonan (RG-I and RG-II) with very complex structures. The backbone of RG-I consists of alternating α-1,4-linked D- galacturonic acid and α-1,2-linked L-rhamnose residues, and its side chains contain D-galactose and/or L-arabinose chains, sometimes with terminal ferulic acid residues. RG-II has an HG backbone of at least eight (and most probably more) 1,4-linked α-D-galacturonic acid decorated with four side branches consisting of 12 glycosyl residues linked together by over 20 different linkages. All RG-II structures contain D-galacturonic acid, D-glucuronic acid, L-rhamnose, D-galactose, L-arabinose, and L-fucose [34].

Due to the highly complex structure of pectin, the degradation of pectin requires a broad set of enzymatic activities (Table 1). These enzymes are classified into more than 20 CAZy families (Figure 2), mainly including pectin hydrolases (GH2, GH5_16, GH28, GH35, GH43, GH51, GH54, GH78, GH88, GH93 and GH105), pectin lyases (PL1, PL3, PL4, PL9 and PL11) [35], and pectin methyl esterases (CE8), and pectin and rhamnogalacturonan acetyl esterases (CE12 [36]).

### *2.4 Storage polysaccharides degradation*

The main forms of plant storage polysaccharides are starch and inulin, although also a variety of gums with different structures can be found. Starch mainly consists of α-1,4-linked polymer (amylose) of D-glucose residues which can be branched at α-1,6-linked points (amylopectin) [37-39]. Starch is degraded by the enzymatic actions of α-glucosidase (AGD), amylo-α-1,6-glucosidase (AMG), α-amylase (AMY), glucoamylase (GLA) and LPMOs (Table 1).

Inulin is found in the roots and rhizomes of many plants and consists of a branched β-2,1-linked chain of D-fructose with a terminal D-glucose residue [40, 41]. Although GH32 is the only CAZy family involved in inulin degradation, it has three enzymatic activities: endo-inulinase (INU), exo-

inulinase (INX) and invertase (INV). This polysaccharide is hydrolyzed by endo- and exo-inulinases to D-fructose and short fructo-oligosaccharides [42, 43], while these oligosaccharides and sucrose are broken down into D-glucose and/or D-fructose by the action of fructofuranosidase (invertase) [44]. Previous research has shown that *Aspergillus* and *Penicillium* species are the most prevalent filamentous fungi to produce inulinase. Several fungi have been verified to produce endo-inulinase as well as exo-inulinase, e.g. *Aspergillus niger* [42], *Aspergillus ficuum* [45], *Chrysosporium pannorum* [46], *Penicillium subrubescens* [47] and *Penicillium purpurogenum* [48].

## 3. Fungal sugar metabolism related to plant biomass utilization

After fungal extracellular enzymatic decomposition, the complex plant polysaccharides are degraded into mono- or short oligomers that can be taken up by the fungus and then be further converted by a variety of intracellular metabolic pathways (Figure 3). Plant biomass constitutive monomeric components include D-glucose, D-fructose, D-galactose, D-mannose, L-rhamnose, D-xylose, L-arabinose, D-galacturonic acid and D-glucuronic acid. When these monomers are released from plant polysaccharides, they will be taken up by fungi and enter a variety of specific sugar metabolic pathways.

D-Glucose and D-fructose are energetically favored monosaccharides because they are easily catabolized. D-fructose and D-glucose are phosphorylated into fructose 6-phosphate and glucose 6-phosphate, respectively, before entering glycolysis. D-xylose and L-arabinose are both converted through the pentose catabolic pathway (PCP) to form D-xylulose-5-phosphate, which subsequently enters the pentose phosphate pathway (PPP) [49]. Other sugars, like D-galactose, D-mannose, D-galacturonic acid, L-rhamnose and D-gluconic acid are also not able to enter glycolysis or the PPP directly, but require additional sugar-specific metabolic pathways [50, 51] for their conversion.

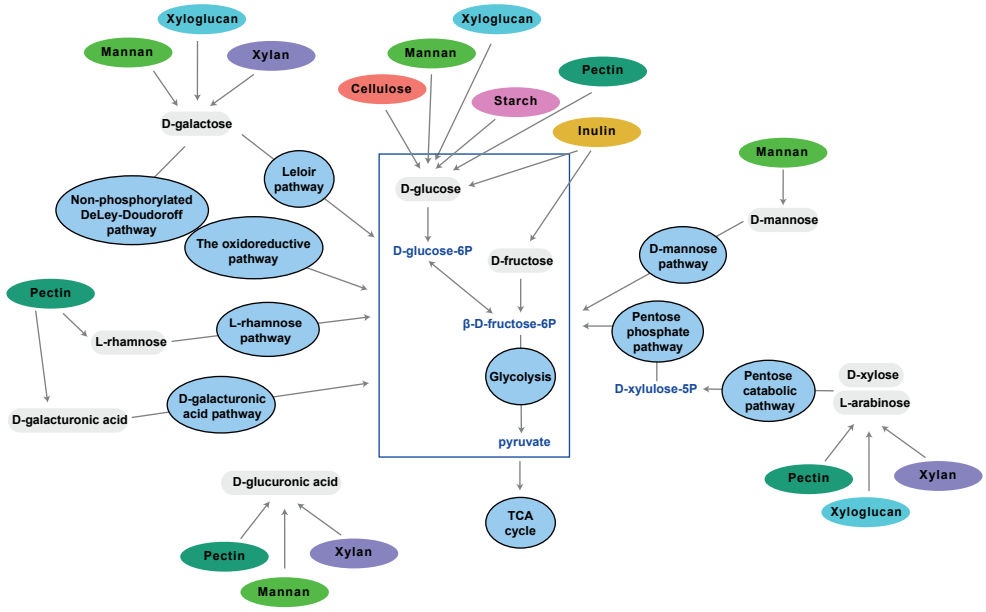


**Figure 3. Fungal sugar metabolism related to the conversion of plant polysaccharides**.

The disaccharides maltose (two α-1,4-linked D-glucose residues) and sucrose (α-1,2 linked D-glucose and D-fructose residues) are components of starch and inulin respectively. They enter glycolysis at different steps according to their monosaccharide composition. In short, fungal central carbon catabolism is a complex network that involves many pathways.

### *3.1 Catabolism of D-glucose and D-fructose through glycolysis*

As mentioned above, D-glucose and D-fructose are the most common constitutive monomers in plant biomass, and converted through glycolysis [52]. The former is phosphorylated to glucose-6-phosphate by glucokinase (GLK, EC 2.7.1.2) [53] and hexokinase (HXK, EC 2.7.1.1) [54], and the latter is converted to D-fructose 6-phosphate [54, 64] by through hexokinase (HXK, EC 2.7.1.4). Afterwards, Glucose-6-phosphate can enter PPP or can be converted to fructose-6-phosphate by glucose 6-phosphate isomerase (PGI, EC 5.3.1.9) [55] to enter glycolysis.

### *3.2 D-Galacturonic acid metabolism*

D-Galacturonic acid is the main constituent of pectin, and it is catabolized through a non-phosphorylating pathway [56]. In the first steps. D-galacturonic acid is reduced to aldehyde-L-galactonate by D-galacturonic acid reductase, which was named as GaaA in *Aspergillus* [57, 58] but GAR1 in *T. reesei* [59]. In the following reaction, aldehyde-L-galactonate is dehydrated to 2-keto-3-deoxy-L-galactonate by the L-galactonate dehydratase (GaaB, EC 4.2.1.146). And then the 2-keto-3-deoxy-L-galactonate is split by 2-keto-3-deoxy-L-galactonate aldolase (GaaC, EC 4.1.2.54) into pyruvate and L-glyceraldehyde. The last step of this pathway requires NADPH as a cofactor and is catalyzed by a glyceraldehyde reductase (GaaD, 1.1.1.372) converting L-glyceraldehyde to glycerol, which enters the glycerol catabolism [60].

### *3.3 D-Mannose metabolism*

D-Mannose forms the backbone of polysaccharide mannan or galactomannan [5]. The catabolism of D-mannose is closely linked to glycolysis. The first enzyme, mannokinase (EC 2.7.1.7) phosphorylates D-mannose to D-mannose-6-phosphate, which is then converted to D-fructose 6-phosphate by mannose-6-phosphate isomerase (PMI, EC 5.3.1.8). D-fructose 6-phosphate then can be fed into glycolysis. In *A. niger*, this first reaction was induced by the same hexokinase [54] that phosphorylates D-glucose into D-glucose 6-phosphate, so it recruited the same gene for both pathways [50]. However, in other studied species, the genes assigned to this activity include not only the genes for the first step in glycolysis, but also the genes involved in the phosphorylation of D-fructose (EC 2.7.1.4) [61]. D-mannose-6-phosphate is also converted to GDP-α-D-mannose through sequential actions of phosphomannomutase (PMM, EC 5.4.2.8) and mannose-1-phosphate guanylyltransferase (MGT, EC 2.7.7.13).

### *3.4 D-Galactose metabolism*

Catabolism of D-galactose contains three (sub-)pathways: the Leloir pathway,

the oxidoreductive pathway [62], and the non-phosphorylated De Ley-Doudoroff pathway (Figure 3).

The Leloir pathway refers to the pathway that converts D-galactose into D-glucose-6-phosphate. [43, 63]. In this pathway, D-galactose is phosphorylated to D-galactose-1-phosphate by galactokinase (GalK, EC 2.7.1.6), and D-galactose-1-phosphate is further converted to D-glucose-1-phosphate by D-galactose-1-phosphate uridylyl transferase (GalT, EC 2.7.7.12). Phosphoglucomutase (PgmB, EC 5.4.2.2) catalyzes the conversion of D-glucose-1-phosphate to D-glucose-6-phosphate.

In the oxidoreductive pathway, β-D-galactose is converted into D-fructose by a series of reductive and oxidative steps [62]. The enzymes involved in this pathway differ between different species [61]. In *T. reesei* and *A. niger*, D-galactose is sequentially converted to D-galactitol, L-xylo-3-hexulose and D-sorbitol [63, 64]. However, L-sorbose was identified as an intermediate product in the conversion between D-galactitol and D-sorbitol in *A. nidulans* [65-68]. Following the conversion of D-sorbitol to D-fructose, D-fructose is then phosphorylated by fructokinase (EC 2.7.1.4) to D-fructose 6-phosphate, which enters glycolysis.

In the non-phosphorylated De Ley-Doudoroff pathway, D-galactose is converted into glyceraldehyde and pyruvate via the intermediates D-galactono-1,4-lactone, D-galactonate, and 2-dehydro-3-deoxy-D-galactonate galactitol. Only the enzymatic activities of the last step of this pathway have been identified, while the genes encoding the enzymes are as yet unknown [69].

### *3.5 Catabolism of D-xylose and L-arabinose through pentose catabolic pathway (PCP)*

L-Arabinose and D-xylose are the most abundant pentoses found in hemicelluloses (such as (arabino)xylan and xyloglucan), and pectin of the plant cell wall [65]. In general, L-arabinose and D-xylose are metabolized by fungi through the pentose catabolic pathway (PCP). The final metabolite,

D-xylulose-5-phosphate, then enters the PPP.

L-arabinose catabolism initiates with the reduction of L-arabinose to L-arabitol through the enzymatic reaction (EC 1.1.1.21) catalyzed by NADPH-dependent L-arabinose reductase (LarA) and D-xylose reductase (XyrA/XyrB). L-arabitol is then oxidized to L-xylulose by L-arabitol-4-dehydrogenase (LadA, EC 1.1.1.12), xylitol dehydrogenase (XdhA), and sorbitol dehydrogenase (SdhA) (EC 1.1.1.12). Subsequently, L-xylulose is reduced to xylitol by two L-xylulose reductases (EC 1.1.1.10): LxrA [70] and LxrB (Terebieniec et al, unpublished data). The oxidation of xylitol into D-xylulose was catalyzed by the same enzymes related to the conversion of L-arabitol to L-xylulose [71].

The L-arabinose and D-xylose pathways share the last two enzymatic steps. The conversion of D-xylose into xylitol is catalyzed by the same enzymes (LarA/XyrA/XyrB, EC 1.1.1.307) involved in the reduction of L-arabinose. Xylitol is further converted into D-xylulose and D-xylulose 5-phosphate sequentially, as described in L-arabinose metabolism [61].

### *3.6 L-Rhamnose metabolism*

L-Rhamnose is enriched in some fractions of plant biomass, such as pectin. The metabolism of L-rhamnose includes four steps: L-rhamnose is first oxidized to L-rhamnose-1,4-lactone by an NADH-dependent L-rhamnose-1-dehydrogenase (LraA, EC 1.1.1.173). L-Rhamnose-1,4-lactone is further metabolized to L-rhamnonate and subsequently to 2-dehydro-3-deoxy-l-rhamnonate in two consecutive reactions catalyzed by L-rhamnonic acid lactonase (LrlA, EC 3.1.1.65) and L-rhamnonate dehydratase (LrdA, EC 4.2.1.90), respectively. Finally, 2-dehydro-3-deoxy-l-rhamnonate is cleaved into pyruvate and L-lactaldehyde by a 2-keto-3-deoxy-L-rhamnonate aldolase (LkaA, EC 4.1.2.53) [61, 72, 73].

## 4. Bioinformatics approaches and omics data for fungal biotechnology

The advancement of high-throughput "omics" technologies, including genomics, transcriptomics, proteomics, and metabolomics, have significantly improved our understanding of the complex plant polysaccharide degrading machinery of individual filamentous fungi, such as the identification of key genes involved in sugar metabolic pathways, and associated sugar transporters and transcriptional regulators. These advancements hold great promise for the development of more efficient and sustainable strategies for biomass conversion, biofuel production, and other biotechnological applications.

In the genome era, as technology advanced to support inexpensive high-throughput genome sequencing, the repositories of fungal genome sequences have been continuously expanding [74]. Thus, the availability of fungal genomes had a tremendous impact on fungal biology as a whole and also on research in plant biomass degradation by fungi. According to the latest report of the Joint Genome Institute (JGI), MycoCosm (https://mycocosm.jgi.doe.gov/) currently includes over 2000 fungal genomes [75]. Efficiently exploring these genomes allows the scientific community to address challenges related to energy and the environment. In addition, the emergence of large sets of fungal transcriptomic, proteomic and metabolomic data has enabled the use of novel methodologies and improved our understanding of the complex molecular mechanisms employed by filamentous fungi for plant polysaccharide degradation. It has enabled the identification of novel genes and enzymes, elucidation of regulatory networks, and characterization of metabolic pathways involved in this process [76-78]. For example, in a recent study of *Neurospora crassa*, transcriptome analysis and DNA affinity purification sequencing (DAP-seq) was integrated to identify regulatory factors and direct targets, respectively, which provides a more complete understanding of cross talk between transcription factors and their target genes [79]. In addition, multi-omics analyses (including transcriptomics,

proteomics and metabolomics) were employed in the study of lignin metabolism in two white-rot fungi, *Trametes versicolor* and *Gelatoporia* (*Ceriporiopsis*) *subvermispora*. The results demonstrate that *T. versicolor* and *G. subvermispora*, in addition to exhibiting different lignocellulose degradation patterns, also likely harbor different aromatic catabolic pathways and/or regulatory mechanisms [80]. Another interesting study is about using a machine learning based approach to predict novel pectinolytic enzymes in *A. niger*, which integrated a diverse range of heterogeneous (post-) genomics data [81]. Therefore, the integration of multi-omics data enables a holistic and systems-level perspective, providing a deeper understanding of complex biological processes. However, the application of this to evaluate differences in the plant biomass degrading approach has only been occasionally done and is typically not very in depth. In this thesis I aim to unearth this diversity of plant biomass conversion approaches using integrated multi-omics, and by comparing the response to pure monosaccharides and crude plant biomass substrates.

In **Chapter 2**, we have comprehensively reviewed the basic bioinformatics applications in fungal research, especially focusing on describing the bioinformatics tools and methods applied in fungal genomics, transcriptomics, proteomics and metabolomics studies, as well as the integrative analysis of multiple omics datasets.

## 5. Aim and outline of this thesis

The aim of this thesis is to use a combination of bioinformatic and omics approaches to obtain a deeper understanding of the molecular mechanisms involved in plant biomass conversion by fungi, and the differences in these approaches across the fungal kingdom.

In this thesis, we systematically evaluated the diversity of the sugar metabolism and CAZymes production of six diverse fungi. Multiple omics data of these fungi grown on diverse monosaccharides and crude biomass conditions have

been integrated and comparatively analyzed to reveal the species- and carbon source- specific molecular response. Six taxonomically distant fungi were selected as the target species, including four Ascomycete species (*Aspergillus niger*, *Aspergillus nidulans*, and *Penicillium subrubescens* belonging to order of Eurotiales, and *Trichoderma reesei* from the Sordariomycetes) and two Basidiomycetes (*Phanerochaete chrysosporium* and *Dichomitus squalens*). Among them, *Trichoderma* and *Aspergillus* are widely used industrial workhorse for production enzymes and bio-products [1, 82, 83]. *P. subrubescens* has been demonstrated as a promising species for producing plant biomass degrading enzymes [84, 85]. Two model white-rot Basidiomycete, *P. chrysosporium* and *D. squalens* were selected because of their strong wood-decaying ability, especially degrading the aromatic polymer lignin [86, 87]. In this thesis, I evaluated the extent of diversity among these six fungi with respect to their genetic response to a diverse set of plant-derived monosaccharides and two crude plant biomass. I especially examined the diversity of sugar metabolism and plant biomass degrading CAZymes based on comparative (functional-)genomics analysis. Results described in this thesis provide deeper insights into the diversity of fungal plant biomass conversion across the fungal kingdom, which will facilitate future metabolic engineering of industrial fungi species and development of more efficient enzymatic cocktails towards more efficient utilization of plant biomass for sustainable production of biofuel, biochemicals and other high value biomolecules.

**Chapter 2** summarizes the commonly applied bioinformatics approaches for current fungal research. This chapter provides a general overview on the basic operations and new developments of bioinformatics applications in fungal research, with a particular emphasis on the bioinformatics tools and methods applied in fungal genomics, transcriptomics, proteomics and metabolomics studies as well as the integrative analysis of multiple omics datasets. Additionally, this chapter highlighted several studies to show the

importance of bioinformatic analysis of fungal omics data in advancing our understanding of fungal physiology and evolutionary diversity.

Sugar catabolism by fungi [51] is directly linked to their ability to convert plant biomass. A detailed insight into sugar catabolism of different species is highly relevant for our understanding of the role of fungi in their natural environment as well as for applying the sugar catabolic pathways for the production of valued and/or novel biochemicals. In the group, my previous colleagues Aguilar-Pontes, M.V., et al., have established the first genome-scale metabolic model for primary carbon metabolism in the laboratory model fungus and industrial workhorse fungus, *A. niger* [50], and this model has a high potential to become a reference for other fungi. In **Chapter 3**, we extrapolated the sugar metabolic model of *A. niger* to five fungi with increasing taxonomic distance to *A. niger* (*A. nidulans*, *P. subrubescens*, *T. reesei*, *P. chrysosporium* and *D. squalens*) to explore to what extent sugar catabolism differs across the fungal kingdom. We provided a detailed metabolic map of these six evolutionarily diverse fungi, and our results revealed that the sugar catabolic pathways were highly conserved in the other two studied Eurotiomycetes (*A. nidulans* and *P. subrubescens*), while the level of diversity was significantly increased between the Sordariomycete *T. reesei* and *A. niger*, and even more so for the Basidiomycota species. In addition, we performed transcriptomics, proteomics and metabolomics analyses, as well as growth profiles on nine related monosaccharides, which further confirmed the sugar metabolic diversity across different fungi.

Filamentous fungi produce extensive sets of extracellular Carbohydrate-Active Enzymes (CAZymes) to be able to degrade complex polysaccharide mixtures in plant biomass to metabolizable small sugars [1]. In **Chapter 4**, we systematically compared the genomic repertoires and transcriptome profile of plant biomass degradation related CAZy genes across six evolutionarily diverse fungi (as described in **Chapter 3**). We discussed in more detail about the diversity of genomic potential with respect to degradation of each specific

polysaccharide. In addition, we comparatively analyzed the expression of CAZymes encoding genes during fungi growth on nine common plant-derived monosaccharides. The results showed noticeable diversity at the genomic and transcriptomic level, suggesting that even in closely related species, their degrading strategies differed markedly. Furthermore, we investigated the possible relationship between the (post-)genomic profile of CAZymes and the growth profiling on 18 plant biomass-related carbon sources, which revealed only limited correlations between the omics and growth phenotypes.

To further investigate the diversity of different fungi with respect to their molecular mechanisms in converting plant biomass, we comparatively analyzed the transcriptome profile of five fungi grown on two crude plant biomass in **Chapter 5**. The results revealed that the expression profiles of sugar metabolic and plant biomass degrading CAZy genes displayed strong time-, substrate- and species-specific differences. The combined analysis of crude plant biomass and monosaccharide (from **Chapter 4**) data suggests complex substrate preference and regulatory networks that differ among the studied fungi.

Finally, the results of each chapter of this thesis are summarized and discussed in **Chapter 6**, including suggesting for future research and the overall relevance of the work presented in this thesis.

## Acknowledgements

JL was supported by the China Scholarship Council (CSC 201909110079).

## References

1. de Vries, R.P. and J. Visser, *Aspergillus enzymes involved in degradation of plant cell wall polysaccharides.* **Microbiology and Molecular Biology Reviews**, 2001. 65(4): p. 497-522.
2. Meyer, V., et al., *Growing a circular economy with fungal biotechnology: a white paper.* **Fungal Biology and Biotechnology**, 2020. 7(1): p. 1-23.
3. Hyde, K.D., et al., *The amazing potential of fungi: 50 ways we can exploit fungi industrially.* **Fungal Diversity**, 2019. 97: p. 1-136.

4. Somerville, C., et al., *Feedstocks for lignocellulosic biofuels.* **Science**, 2010. 329(5993): p. 790-792.
5. Culleton, H., V. McKie, and R.P. de Vries, *Physiological and molecular aspects of degradation of plant polysaccharides by fungi: what have we learned from Aspergillus?* **Biotechnology Journal**, 2013. 8(8): p. 884-894.
6. Cantarel, B.L., et al., *The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics.* **Nucleic Acids Research**, 2009. 37(suppl_1): p. D233-D238.
7. Lombard, V., et al., *The carbohydrate-active enzymes database (CAZy) in 2013.* **Nucleic Acids Research**, 2014. 42(D1): p. D490-D495.
8. Biely, P., et al., *Inverting character of α-glucuronidase A from Aspergillus tubingensis.* **Biochimica et Biophysica Acta (BBA)-General Subjects**, 2000. 1474(3): p. 360-364.
9. Mewis, K., et al., *Dividing the large glycoside hydrolase family 43 into subfamilies: a motivation for detailed enzyme characterization.* **Applied and Environmental Microbiology**, 2016. 82(6): p. 1686-1692.
10. Zhang, X.J., et al., *Contributions and characteristics of two bifunctional GH43 β-xylosidase/α-L-arabinofuranosidases with different structures on the xylan degradation of Paenibacillus physcomitrellae strain XB.* **Microbiological Research**, 2021. 253: p. 126886.
11. Coutinho, P.M., et al., *Post-genomic insights into the plant polysaccharide degradation potential of Aspergillus nidulans and comparison to Aspergillus niger and Aspergillus oryzae.* **Fungal Genetics and Biology**, 2009. 46(1): p. S161-S169.
12. Villarreal, F., N. Stocchi, and A. Ten Have, *Functional classification and characterization of the fungal Glycoside Hydrolase 28 protein family.* **Journal of Fungi**, 2022. 8(3): p. 217.
13. Li, X., et al., *Glycoside Hydrolase family 30 harbors fungal subfamilies with distinct polysaccharide specificities.* **New Biotechnology**, 2022. 67: p. 32-41.
14. Helbert, W., et al., *Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space.* **Proceedings of the National Academy of Sciences**, 2019. 116(13): p. 6063-6068.
15. van den Brink, J. and R.P. de Vries, *Fungal enzyme sets for plant polysaccharide degradation.* **Applied Microbiology and Biotechnology**, 2011. 91(6): p. 1477-1492.
16. Bornscheuer, U., K. Buchholz, and J. Seibel, *Enzymatic degradation of (ligno) cellulose.* **Angewandte Chemie International Edition**, 2014. 53(41): p. 10876-10893.
17. Ezeilo, U.R., et al., *Enzymatic breakdown of lignocellulosic biomass: the role of glycosyl hydrolases and lytic polysaccharide monooxygenases.* **Biotechnology &**

**Biotechnological Equipment**, 2017. 31(4): p. 647-662.

18. Filiatrault-Chastel, C., et al., *AA16, a new lytic polysaccharide monooxygenase family identified in fungal secretomes.* **Biotechnology for Biofuels**, 2019. 12(1): p. 1-15.
19. Andlar, M., et al., *Lignocellulose degradation: An overview of fungi and fungal enzymes involved in lignocellulose degradation.* **Engineering in Life Sciences**, 2018. 18(11): p. 768-778.
20. Filiatrault-Chastel, C., et al., *From fungal secretomes to enzymes cocktails: The path forward to bioeconomy.* **Biotechnology Advances**, 2021. 52: p. 107833.
21. Dimarogona, M., E. Topakas, and P. Christakopoulos, *Cellulose degradation by oxidative enzymes.* **Computational and Structural Biotechnology Journal**, 2012. 2(3): p. e201209015.
22. Kracher, D. and R. Ludwig, *Cellobiose dehydrogenase: An essential enzyme for lignocellulose degradation in nature–A review/Cellobiosedehydrogenase: Ein essentielles Enzym für den Lignozelluloseabbau in der Natur–Eine Übersicht.* **Die Bodenkultur: Journal of Land Management, Food and Environment**, 2016. 67(3): p. 145-163.
23. Langston, J.A., et al., *Oxidoreductive cellulose depolymerization by the enzymes cellobiose dehydrogenase and glycoside hydrolase 61.* **Applied and Environmental Microbiology**, 2011. 77(19): p. 7007-7015.
24. Scheller, H.V. and P. Ulvskov, *Hemicelluloses.* **Annual Review of Plant Biology**, 2010. 61: p. 263-289.
25. Rytioja, J., et al., *Plant-polysaccharide-degrading enzymes from basidiomycetes.* **Microbiology and Molecular Biology Reviews**, 2014. 78(4): p. 614-649.
26. Collins, T., C. Gerday, and G. Feller, *Xylanases, xylanase families and extremophilic xylanases.* **FEMS Microbiology Reviews**, 2005. 29(1): p. 3-23.
27. Li, X., et al., *Fungal xylanolytic enzymes: Diversity and applications.* **Bioresource Technology**, 2022. 344: p. 126290.
28. Venegas, F.A., et al., *Carbohydrate esterase family 16 contains fungal hemicellulose acetyl esterases (HAEs) with varying specificity.* **New Biotechnology**, 2022. 70: p. 28-38.
29. Schultink, A., et al., *Structural diversity and function of xyloglucan sidechain substituents.* **Plants**, 2014. 3(4): p. 526-542.
30. Damásio, A.R., et al., *Functional characterization and oligomerization of a recombinant xyloglucan-specific endo-β-1, 4-glucanase (GH12) from Aspergillus niveus.* **Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics**, 2012. 1824(3): p. 461-467.
31. Sun, P., et al., *Fungal glycoside hydrolase family 44 xyloglucanases are restricted to the phylum Basidiomycota and show a distinct xyloglucan cleavage pattern.*

**Iscience**, 2022. 25(1): p. 103666.

32. Matsuzawa, T., et al., *Identification and characterization of two xyloglucan-specific endo-1, 4-glucanases in Aspergillus oryzae.* **Applied Microbiology and Biotechnology**, 2020. 104(20): p. 8761-8773.
33. Ademark, P., et al., *Cloning and characterization of Aspergillus niger genes encoding an α-galactosidase and a β-mannosidase involved in galactomannan degradation.* **European Journal of Biochemistry**, 2001. 268(10): p. 2982-2990.
34. Caffall, K.H. and D. Mohnen, *The structure, function, and biosynthesis of plant cell wall pectic polysaccharides.* **Carbohydrate Research**, 2009. 344(14): p. 1879-1900.
35. Benoit, I., et al., *Degradation of different pectins by fungi: correlations and contrasts between the pectinolytic enzyme sets identified in genomes and the growth on pectins of different origin.* **BMC Genomics**, 2012. 13(1): p. 1-11.
36. Mølgaard, A., S. Kauppinen, and S. Larsen, *Rhamnogalacturonan acetylesterase elucidates the structure and function of a new family of hydrolases.* **Structure**, 2000. 8(4): p. 373-383.
37. Mischnick, P. and D. Momcilovic, *Chemical structure analysis of starch and cellulose derivatives.* **Advances in Carbohydrate Chemistry and Biochemistry**, 2010. 64: p. 117-210.
38. Gamage, A., et al., *Applications of starch biopolymers for a sustainable modern agriculture.* **Sustainability**, 2022. 14(10): p. 6085.
39. Taggart, P., *Starch as an ingredient: manufacture and applications.* **Starch in food: Structure, Function and Applications**, 2004: p. 363-392.
40. Pavis, N., et al., *Structure of fructans in roots and leaf tissues of Lolium perenne.* **New Phytologist**, 2001. 150(1): p. 83-95.
41. Roberfroid, M.B., *Introducing inulin-type fructans.* **British Journal of Nutrition**, 2005. 93(S1): p. S13-S25.
42. Derycke, D.G. and E.J. Vandamme, *Production and properties of Aspergillus niger inulinase.* **Journal of Chemical Technology and Biotechnology. Biotechnology**, 1984. 34(1): p. 45-51.
43. Chen, H.-Q., et al., *Purification and characterisation of exo-and endo-inulinase from Aspergillus ficuum JNSP5-06.* **Food Chemistry**, 2009. 115(4): p. 1206-1212.
44. Wallis, G., F. Hemming, and J. Peberdy, *Secretion of two β-fructofuranosidases by Aspergillus niger growing in sucrose.* **Archives of Biochemistry and Biophysics**, 1997. 345(2): p. 214-222.
45. Ettalibi, M. and J.C. Baratti, *Purification, properties and comparison of invertase, exoinulinases and endoinulinases of Aspergillus ficuum.* **Applied Microbiology and Biotechnology**, 1987. 26: p. 13-20.
46. Xiao, R., M. Tanida, and S. Takao, *Purification and some properties of endoinulinase*

*from Chrysosporium pannorum.* **Journal of Fermentation and Bioengineering**, 1989. 67(4): p. 244-248.

47. Mansouri, S., et al., *Penicillium subrubescens, a new species efficiently producing inulinase.* **Antonie van Leeuwenhoek**, 2013. 103: p. 1343-1357.
48. Onodera, S., et al., *Molecular cloning and nucleotide sequences of cDNA and gene encoding endo-inulinase from Penicillium purpurogenum.* **Bioscience, Biotechnology, and Biochemistry**, 1996. 60(11): p. 1780-1785.
49. Chroumpi, T., et al., *Revisiting a 'simple' fungal metabolic pathway reveals redundancy, complexity and diversity.* **Microbial Biotechnology**, 2021. 14: p. 2525-2537.
50. Aguilar-Pontes, M.V., et al., *The gold-standard genome of Aspergillus niger NRRL 3 enables a detailed view of the diversity of sugar catabolism in fungi.* **Studies in Mycology**, 2018. 91: p. 61-78.
51. Khosravi, C., et al., *Sugar catabolism in Aspergillus and other fungi related to the utilization of plant biomass.* **Advances in Applied Microbiology**, 2015. 90: p. 1-28.
52. Romano, A. and T. Conway, *Evolution of carbohydrate metabolic pathways.* **Research in Microbiology**, 1996. 147(6-7): p. 448-455.
53. Panneman, H., et al., *Cloning and biochemical characterisation of an Aspergillus niger glucokinase: Evidence for the presence of separate glucokinase and hexokinase enzymes.* **European Journal of Biochemistry**, 1996. 240(3): p. 518-525.
54. Panneman, H., et al., *Cloning and biochemical characterisation of Aspergillus niger hexokinase: The enzyme is strongly inhibited by physiological concentrations of trehalose 6-phosphate.* **European Journal of Biochemistry**, 1998. 258(1): p. 223-232.
55. Ruijter, G.J. and J. Visser, *Characterization of Aspergillus niger phosphoglucose isomerase. Use for quantitative determination of erythrose 4-phosphate.* **Biochimie**, 1999. 81(3): p. 267-272.
56. Richard, P. and S. Hilditch, *D-galacturonic acid catabolism in microorganisms and its biotechnological relevance.* **Applied Microbiology and Biotechnology**, 2009. 82(4): p. 597-604.
57. Martens-Uzunova, E.S. and P.J. Schaap, *An evolutionary conserved D-galacturonic acid metabolic pathway operates across filamentous fungi capable of pectin degradation.* **Fungal Genetics and Biology**, 2008. 45(11): p. 1449-1457.
58. Hilditch, S., *Identification of the fungal catabolic D-galacturonate pathway*. 2010, Espoo: VTT Technical Research Centre of Finland.
59. Kuorelahti, S., et al., *Identification in the mold Hypocrea jecorina of the first fungal D-galacturonic acid reductase.* **Biochemistry**, 2005. 44(33): p. 11234-11240.
60. Alazi, E., et al., *The pathway intermediate 2-keto-3-deoxy-L-galactonate mediates the induction of genes involved in D-galacturonic acid utilization in Aspergillus*

*niger.* **FEBS Letters**, 2017. 591(10): p. 1408-1418.

61. Li, J., et al., *The sugar metabolic model of Aspergillus niger can only be reliably transferred to fungi of its phylum.* **Journal of Fungi**, 2022. 8(12): p. 1315.
62. Mojzita, D., et al., *L-xylo-3-hexulose reductase is the missing link in the oxidoreductive pathway for D-galactose catabolism in filamentous fungi.* **Journal of Biological Chemistry**, 2012. 287(31): p. 26010-26018.
63. Pail, M., et al., *The metabolic role and evolution of L-arabinitol 4-dehydrogenase of Hypocrea jecorina.* **European Journal of Biochemistry**, 2004. 271(10): p. 1864-1872.
64. Mojzita, D., et al., *Identification of the galactitol dehydrogenase, LadB, that is part of the oxido-reductive D-galactose catabolic pathway in Aspergillus niger.* **Fungal Genetics and Biology**, 2012. 49(2): p. 152-159.
65. Seiboth, B. and B. Metz, *Fungal arabinan and L-arabinose metabolism.* **Applied Microbiology and Biotechnology**, 2011. 89(6): p. 1665-1673.
66. Fekete, E., et al., *The alternative D-galactose degrading pathway of Aspergillus nidulans proceeds via L-sorbose.* **Archives of Microbiology**, 2004. 181(1): p. 35-44.
67. Orosz, A., et al., *Metabolism of D-galactose is dispensable for the induction of the beta-galactosidase (bgaD) and lactose permease (lacpA) genes in Aspergillus nidulans.* **FEMS Microbiology Letters**, 2014. 359(1): p. 19-25.
68. Kowalczyk, J.E., et al., *Genetic interaction of Aspergillus nidulans galR, xlnR and araR in regulating D-galactose and L-arabinose release and catabolism gene expression.* **PLoS One**, 2015. 10(11): p. e0143200.
69. Elshafei, A.M. and O.M. Abdel-Fatah, *Evidence for a non-phosphorylated route of galactose breakdown in cell-free extracts of Aspergillus niger.* **Enzyme and Microbial Technology**, 2001. 29(1): p. 76-83.
70. Mojzita, D., et al., *The 'true'L-xylulose reductase of filamentous fungi identified in Aspergillus niger.* **FEBS Letters**, 2010. 584(16): p. 3540-3544.
71. Chroumpi, T., et al., *Revisiting a 'simple'fungal metabolic pathway reveals redundancy, complexity and diversity.* **Microbial Biotechnology**, 2021. 14(6): p. 2525-2537.
72. Khosravi, C., et al., *In vivo functional analysis of L-rhamnose metabolic pathway in Aspergillus niger: a tool to identify the potential inducer of RhaR.* **BMC Microbiology**, 2017. 17(1): p. 1-12.
73. Chroumpi, T., et al., *Identification of a gene encoding the last step of the L-rhamnose catabolic pathway in Aspergillus niger revealed the inducer of the pathway regulator.* **Microbiological Research**, 2020. 234: p. 126426.
74. Stajich, J.E., *Fungal genomes and insights into the evolution of the kingdom.* **Microbiology Spectrum**, 2017. 5(4): p. 5.4. 15.

75. Ahrendt, S.R., et al., *MycoCosm, the JGI's fungal genome portal for comparative genomic and multiomics data analyses*, in **Microbial Environmental Genomics (MEG)**. 2022, Springer. p. 271-291.
76. Rokas, A., et al., *What can comparative genomics tell us about species concepts in the genus Aspergillus?* **Studies in Mycology**, 2007. 59(1): p. 11-17.
77. Moran, G.P., D.C. Coleman, and D.J. Sullivan, *Comparative genomics and the evolution of pathogenicity in human pathogenic fungi.* **Eukaryotic Cell**, 2011. 10(1): p. 34-42.
78. Peng, M., et al., *Comparative analysis of basidiomycete transcriptomes reveals a core set of expressed genes encoding plant biomass degrading enzymes.* **Fungal Genetics and Biology**, 2018. 112: p. 40-46.
79. Wu, V.W., et al., *The regulatory and transcriptional landscape associated with carbon utilization in a filamentous fungus.* **Proceedings of the National Academy of Sciences**, 2020. 117(11): p. 6003-6013.
80. Del Cerro, C., et al., *Intracellular pathways for lignin catabolism in white-rot fungi.* **Proceedings of the National Academy of Sciences**, 2021. 118(9): p. e2017381118.
81. Peng, M. and R.P. de Vries, *Machine learning prediction of novel pectinolytic enzymes in Aspergillus niger through integrating heterogeneous (post-) genomics data.* **Microbial Genomics**, 2021. 7(12).
82. Peterson, R. and H. Nevalainen, *Trichoderma reesei RUT-C30–thirty years of strain improvement.* **Microbiology**, 2012. 158(1): p. 58-68.
83. Mäkelä, M.R., N. Donofrio, and R.P. de Vries, *Plant biomass degradation by fungi.* **Fungal Genetics and Biology**, 2014. 72: p. 2-9.
84. Dilokpimol, A., et al., *Penicillium subrubescens adapts its enzyme production to the composition of plant biomass.* **Bioresource Technology**, 2020. 311: p. 123477.
85. Mäkelä, M.R., et al., *Penicillium subrubescens is a promising alternative for Aspergillus niger in enzymatic plant biomass saccharification.* **New Biotechnology**, 2016. 33(6): p. 834-841.
86. Martinez, D., et al., *Genome sequence of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78.* **Nature Biotechnology**, 2004. 22(6): p. 695-700.
87. Kowalczyk, J.E., et al., *The white-rot basidiomycete Dichomitus squalens shows highly specific transcriptional response to lignocellulose-related aromatic compounds.* **Frontiers in Bioengineering and Biotechnology**, 2019. 7: p. 229.

# Chapter 2

# Bioinformatics approaches for fungal biotechnology

This chapter was published in *Encyclopedia of Mycology*
Jiajia Li, Ronald P. de Vries, Mao Peng
Volume 2, Pages 536-554, June 2021
DOI: https://doi.org/10.1016/B978-0-12-819990-9.00012-3

## Abstract

Current biological research has been revolutionized by various “omics” technologies, which can produce vast quantities of data and enable a systematical monitoring of the genome-wide changes of e.g. genes (genomics), mRNA (transcriptomics), proteins (proteomics) and metabolites (metabolomics) in a specific biological sample. Over the past decades, a large number of bioinformatics software and databases have been developed to analyze the omics data flood and interpret their biological meaning. In this chapter, we give an overview on the commonly applied bioinformatics approaches in today’s fungal research, especially focusing on the tools and methods used for the analysis of fungal genomics, transcriptomics, proteomics and metabolomics data.

## 1. Introduction

Bioinformatics analysis plays a crucial role in today's biological research. Being an interdisciplinary branch of the life sciences, bioinformatics integrates computer science and mathematical methods to reveal the biological significance behind the continually increasing biological data. It does this by developing methodology and analysis tools to explore the large volumes of biological data, helping to query, extract, store, organize, systematize, annotate, visualize, mine, and interpret complex data. In this chapter, we aim to give a general overview on the basic operations and new developments of bioinformatics applications in fungal research, especially focusing on describing the bioinformatics tools and methods applied in fungal genomics, transcriptomics, proteomics and metabolomics studies, as well as the integrative analysis of multiple omics datasets. For each aspect, we highlight several examples to demonstrate how various bioinformatics methods have been successfully used to mine fungal omics data and improve our understanding of fungal physiology and evolution, specially emphasizing bioinformatics applications related to fungal plant biomass conversion.

## 2. Current states of fungal genomics analysis

The sequenced genome together with its annotation is the basis for understanding the biological process and evolutionary history of an organism. Due to the wide application of next generation sequencing (NGS) technology, biological research has entered the genome era. More than 1500 fungal genomes have been sequenced and still many more genomes are in progress since the first fungal genome sequence (*Saccharomyces cerevisiae*) was published in 1996 [1]. For instance, a broad range of reference fungal genomes have been sequenced, including the yeast *Schizosaccharomyces pombe* [2], the filamentous saprobes *Neurospora crassa* [3], *Aspergillus nidulans* [4], *Aspergillus niger* [5] and *Trichoderma reesei* [6], the filamentous cereal pathogens *Magnaporthe grisea* [7], *Ustilago maydis* [8] and *Fusarium graminearum* [9], and the white rot basidiomycete *Phanerochaete chrysosporium* [10]. Moreover, the sequenced

genomes together with their annotation information can be easily accessed from a list of the databases dedicated to fungal genomes (Table 1). The rapidly increasing amount of fungal genomics data not only provides us with a better understanding of the fundamental cellular processes, but also benefits a broad range of fungi related applications in human health, agriculture, enzyme biotechnology, bioenergy, and ecological diversity [11].

**Table 1. List of databases dedicated to fungal genomes.**

| Database | URL |
|---|---|
| MycoCosm | https://mycocosm.jgi.doe.gov/mycocosm/home |
| Joint Genome Institute (JGI) | https://genome.jgi.doe.gov/portal/ |
| FungiDB | https://fungidb.org/fungidb/ |
| EnsemblFungi | https://fungi.ensembl.org/index.html |
| *Aspergillus* Genome Database (AspGD) | http://www.aspgd.org/ |
| *Saccharomyces* Genome Database (SGD) | https://www.yeastgenome.org/ |
| *Candida* Genome Database (CGD) | http://www.candidagenome.org/ |

### *2.1 Bioinformatics workflow and tools for genomics analysis*

The advent of NGS technologies is accompanied with the rapid development of numerous bioinformatics methods and tools that were created to handle the vast amount and different types of data generated by different sequencing platforms. A specific set of bioinformatics tools are needed to process data in each different step of genomic analysis. Space and expertise constraints do not allow us to cover all bioinformatics tools deployed in the genomics field and here we mainly focus on tools that are frequently used in the community.

The past decades have witnessed dramatical changes in sequencing technologies. Early efforts of fungal genome sequencing were mainly made by the Sanger sequencing method, such as the genome sequencing of *S. cerevisiae* [1], *T. reesei* [6], *P. chrysosporium* [10] and *Agaricus bisporus* [12]. However, due to the high cost and time-consuming use of the Sanger method, more recent genomes were mainly sequenced by alternative sequencing platforms, called “next generation sequencing (NGS)”. Several NGS sequencing platforms are widely used by researchers, including Roche

454, Illumina, Helicos, ABI SOLiD and newly developed Ion Proton, PacBio and Oxford Nanopore. These sequencing platforms differ substantially in terms of their sequencing mechanism, throughput, output (length of reads, number of sequences), accuracy and cost [13]. The decrease of sequencing cost compared to traditional Sanger sequencing has driven the rapid development of NGS and made it the primary sequencing method. In this chapter, we focus on the bioinformatics analysis of NGS data.



A typical workflow for genome analysis falls into the following steps: (1) quality control and preprocessing of raw reads; (2) alignment (especially for comparative assembly); (3) genome assembly; (4) genome annotation; (5) other advanced analyses based on your research interest, such as phylogenetic analysis and comparative genomics analysis. Typically, the raw reads of NGS data are stored either as text in a Fasta file or with their qualities as a FastQ file. The latter one is the most widely accepted format. Quality control (QC) of raw reads is the first step of the genomic data analysis and several tools are commonly used for this step. For instance, FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) is a powerful QC tool, which can provide a range of useful QC statistics results, such as per-base quality, average read quality, GC content, N content, sequence length distribution, duplication levels, and overrepresented sequences. Read trimming is a preprocess step which aims at removing unwanted information of sequence, including leftover adapter sequences, low-quality bases at reads ends, known contaminants (strings of As/Ts), reads containing too many N-bases along their length, and more. Popular trimming tools include the FASTA-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), Trimmomatic [14], Fastx_trimmer (https://rdrr.io/github/abshah/RADseqR/man/fastx_trimmer.html) and cutadapt [15]. In addition, NGS QC Toolkit [16] integrates various tools for performing multiple tasks, including quality control, trimming, format conversion, and generates corresponding statistics reports. After filtering the reads, the clean reads are aligned to the reference genome during a comparative

assembly. Widely used aligners include Bowtie2 [17], Burrows-Wheeler Aligner (BWA) [18], STAR [19], HISAT2 [20], and StringTie [21]. The alignments results are usually stored as SAM files or its binary compressed implementation BAM files. Alignments can be viewed using user-friendly and freely available software, such as the Interactive Genome Viewer (IGV) [22], the GALAXY server [23] or Genome Browse (http://goldenhelix.com/products/GenomeBrowse).

Genome assembly refers to the process of taking a large number of short DNA sequences and putting them back together to create a representation of the original chromosomes from which the DNA originated. The genome assembly can be broadly categorized into two approaches: *de novo* assembly, for reconstructing genomes without available reference genomes, and comparative assembly, which uses the sequence of a closely related organism as a reference during assembly. Comparative approaches can only be applied to the genomes for which reference sequences are available. Even when a reference genome is available, the sequence can also be reconstructed through *de novo* assembly to improve the genome [24]. In terms of different demands, there are several computational tools commonly used for genome assembly (Table 2).

Alongside the steady increasing performance of NGS technologies in genome sequencing, it still faces several persistent challenges to achieve a fully assembled genome. For example, errors in reads and large repeats in the genome complicate sequence alignment and genome reconstruction. Therefore, it is necessary to statistically assess the quality of an assembly. The quality of genome assembly can be assessed by several criteria, such as per-base error rates, insert size distributions, k-mer distributions and fragment (contig) length distributions. N50 is a widely used measure to assess the contiguity of a genome assembly, which is defined as the sequence length of the shortest contig for which longer and equal length contigs cover at least 50% of the genome. For quality evaluation of assembly results, several

software can be applied, such as GAGE (Genome Assembly Gold-standard Evaluations) [32], QUAST [33], BUSCO [34], SQUAT [35] and GenomeQC [36]. GAGE could evaluate repeat copy numbers, insertions, deletions, and assembly errors in the various assemblies, even the correctness of assemblies if the reference genomes are given. QUAST can evaluate assemblies with or without a reference genome, and usually presents comprehensive metrics, summary tables, reports and plots. BUSCO does a quantitative assessment of genome assembly and annotation completeness. SQUAT tool (https://github.com/luke831215/SQUAT) provides quality assessments for both pre-assembly and post-assembly, and examines the correctness of *de novo* assemblies and the input reads. GenomeQC characterizes both assembly and gene annotation quality, and provides a comprehensive assessment summary.

The next step after genome assembly is to provide functional annotation. The annotation is a bridge connecting the sequence with its possible biological function. Genome annotation includes structural annotation and functional annotation. The former refers to the identification features on a genome assembly such as protein-coding, non-coding RNAs (ncRNAs), repeats and transposable elements, and pseudogenes. There is a variety of programs available for *ab initio* gene predictions including Genscan [37], Genemark.hmm [38], SNAP [39], GlimmerHMM [40], MAKER [41] and AUGUSTUS [42]. To improve the accuracy of gene prediction and annotation, the NGS based transcriptome data is often integrated by many gene prediction tools. In addition, the RepeatMasker (http://www.repeatmasker.org) is routinely used to identify the repeat elements and mask them from the genome in preparation for a more focused gene finding exercise. The TransposonPSI (http://transposonpsi.sourceforge.net/) is often used to assist in the discovery of transposable elements.

**Table 2. Common assemblers for diverse sequencing data.**

| Tools | Description | Availability | Ref |
|---|---|---|---|
| ABySS | paired-end sequence assembler designed for large genome assembly of short reads | https://github.com/bcgsc/abyss | [25] |
| SOAPdenovo | short read assembler | https://www.animalgenome.org/bioinfo/resources/manuals/SOAP.html | [26, 27] |
| ALLPATHS-LG | short read assembler | http://software.broadinstitute.org/allpaths-lg/blog/ | [28] |
| Canu | long read assembler working on both third and fourth generation reads | https://github.com/marbl/canu | [29] |
| Falcon | long read assembler designed for diploid organisms. | https://github.com/PacificBiosciences/FALCON | [30] |
| Velvet | *de novo* genome assembly and short read sequencing alignments | https://www.ebi.ac.uk/~zerbino/velvet/ | [31] |

Functional annotation refers to the computational assignment of functions to the predicted genes. A common approach is using homology-based methods like BLASTp to infer biological function based on top hits in a well-curated and non-redundant database like the manually annotated SWISS-PROT provided at UniProt (https://www.uniprot.org/) and the reference sequence (RefSeq) collection of the NCBI (https://www.ncbi.nlm.nih.gov/refseq/) A widely used protein homology-based gene-modeling tool is GeneWise (https://www.ebi.ac.uk/Tools/psa/genewise/), which combines protein alignment and gene prediction into a single statistical model via a paired hidden Markov model (HMM) [43]. Additional functional characterization includes assigning enzyme activity (EC-number), gene ontology (GO), Pfam database (https://pfam.xfam.org/), and KEGG-pathway. Several tools are commonly used for this analysis like InterProScan (http://www.ebi.ac.uk/interpro/about/interpro/), Blast2GO software (https://www.blast2go.com/), the "hmmsearch" of the HMMER tool (http://hmmer.org/), and Funannotate (https://funannotate.readthedocs.io/en/latest/tutorials.html). Furthermore, the secretion signal peptide and transmembrane characters can be predicted with

the SignalP tool (http://www.cbs.dtu.dk/services/SignalP/), and TMHMM (http://www.cbs.dtu.dk/services/TMHMM/), respectively. Additionally, the comparative fungal genomics platform (CFGP, http://cfgp.riceblast.snu.ac.kr/main.php), is a multifunctional fungal genome data warehouse, which enables online analysis of large set of fungal genomes by integrating 37 common bioinformatics tools and encompasses several specific repositories focusing on specific fungal protein groups such as secretome, transcription factor and cytochrome P450 proteins [44].



The annotation information of non-coding RNA is often derived from a reference genome. If this is not available, the common method is to map non-coding RNA sequences with known members from databases of non-coding RNA. Rfam [45, 46] and INFERNAL [47] serve useful resources for annotating transfer RNA (tRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), and other short non-coding RNA with conserved sequences and secondary structures. NONCODE (http://www.noncode.org) is an integrated knowledgebase dedicated to the complete collection and annotation of non-coding RNAs (ncRNA) [48]. lncRNAdb (https://rnacentral.org/expert-database/lncrnadb) provides annotation of long non-coding RNAs [49].

Another method of functional annotation is the classification of orthologous genes. OrthoMCL [50] and OrthoFinder [51] are commonly used to analyze orthologues. Due to their importance biological function and industrial potential, some genes have been further annotated and classified into specific protein families, such as the carbohydrate-active enzyme annotation (CAZyme) and peptidases. CAZyme predictions are usually performed by running hmmscan from the HMMER tool (http://hmmer.org/) by searching the HMM profiles derived from well-characterized CAZymes [52, 53]. Similar approach could be used to identify the peptidases by using sequences provided on MEROPS website (https://www.ebi.ac.uk/merops) and HMMER tools. The resulting set of genes and functional annotation is typically less

than perfect, and often need manual curation before its release to the public, which requires continuous refining efforts from the scientific community even after the release of the genome.

To make the best use of a genome, the newly assembled genome together with its annotation is uploaded to public databases, which can be accessed and shared by researchers worldwide. The most common databases dedicated to fungal genomes are shown in Table 1. In general, there are two types of databases. Several databases embrace a broader diversity of fungi, such as Mycocosm and FungiDB. Some other database focus on specific fungal species or genera. For examples, the SGD [54] and CGI [55] are specialized databases for *Saccharomycetes* and *Candida*, respectively. AspGD was designed for *Aspergillus* researchers, which provides visualization tools for genomic features and alignments as well as a comparative genomics analysis for the genus *Aspergillus* [56]. Currently the AspGD database is no longer updated, and AspGD data has been integrated into FungiDB (https://fungidb.org/fungidb/). Given their economic and scientific importance, several fungal genomes have received more attention and are selected to be deeply annotated to achieve gold-standard genome resources. Those flagship fungi include *A. niger* [5], *A. nidulans* [4], *T. reesei* [6], and *P. chrysosporium* [10].

### *2.2 Examples of fungal genomic studies*

With abundant and diverse fungal genomes available, comparative genomics has emerged as a powerful approach to identify the link between genotype and phenotype via comparatively analyzing genomic features and their evolution across different species. In the past decade, numerous studies of comparative genomics have greatly enhanced our understanding of the genomic diversity and evolutionary mechanisms that are related to several important aspects of fungal biology, such as plant pathogenicity [57-59], plant biomass degradation [60-63], as well as stress response [64]. The current rapid expansion of genomics data allows even large-scale comparison of genomics at the genera or section level. For example, a recent comparative

analysis of a large set of *Aspergillus* genomes have revealed remarkable genomic diversity and similarities within *Aspergillus* species and part of them could be associated with specific phenotypes [65]. Another comparative genomics study of 23 *Aspergillus* species from section *Flavi* reassessed the phylogenetic relationships of the studied species, and highlighted the genetic and metabolic diversity within section *Flavi* by comparing evolution patterns of CAZymes, secondary metabolite gene clusters and fungal growth profiles [66]. A similar study of comparing inter- and intraspecies variation of 23 *Aspergillus* section *Nigri* also showed high genetic diversity and genome flexibility of section *Nigri* [67]. Similar comparative studies have also been performed for 12 *Trichoderma* species [68] and various sets of basidiomycete fungi. A comparative analysis of 31 basidiomycete genomes, including both white and brown rot basidiomycetes, reconstructed the evolution of lignin decay mechanisms and suggested that the evolution of ligninolytic capabilities was associated with the sharp decrease of carbon burial in geological time [60]. A further comparison of 33 basidiomycete genomes revealed that the current dichotomous classification of white rot and brown rot fungi was not adequate to describe the whole spectrum of wood-decaying basidiomycetes [69]. For instance, *Botryobasidium botryosum* and *Jaapia argillacea* have lost ligninolytic class II peroxidases like brown rot fungi, but possess diverse enzymes acting on crystalline cellulose normally associated with white rot fungi, and thus represent so-called grey rot species. In summary, with the continuous expansion of fungal genomics, comparative genomics analysis have an enormous potential to help us better understand how fungi have adapted to their lifestyles and ecological niches, and to discover new biochemical mechanisms [70].



## 3. Current states of fungal post-genomics analysis

In parallel with the increase in fungal genome sequences, the amount of fungal functional genomic data (transcriptomics, proteomics and metabolomics data) has also increased rapidly. Meanwhile, a broad range of bioinformatics tools

have been developed as a response to the exponential growth in the amount, scale, and diversity of post-genomics data. In the post-genomic era, the focus of research will shift from revealing the genetic information of an organism to the functional biological research.

### *3.1 Transcriptomics*

The transcriptome is the complete set of all RNA molecules in a cell, a population of cells or an organism. There are several major types of RNA inside a cell. The most studied one is the messenger RNA (mRNA) which is translated into proteins. Other types of RNAs serve other important functions, such as circular RNA (circRNA), microRNA (miRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), small interfering RNA (siRNAs), and long non-coding RNA (lncRNA). Here we focus on mRNA related transcriptomics performed with RNA-Seq, which is the most commonly used experimental approach in current fungal studies. We focus on the bioinformatics analysis, and do not address the experimental details.

#### *3.1.1 Bioinformatics workflow and tools for transcriptomics analysis*

Generally speaking, the RNA-Seq downstream analysis can include four essential stages: quality control, alignment, qualification, and differential analysis. Several NGS sequencing platforms are used by researchers, including Roche 454, Illumina, Helicos, ABI SOLiD and PacBio [71]. Among them Illumina is the most commonly used platform [72]. Depending on the research requirement, single-end and/or paired-end reads can be generated. After obtaining a large volume of raw sequence reads, similar as for genomic analysis, quality analysis is needed to remove low quality reads by using tools such as FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), FaQCs (https://github.com/LANL-Bioinformatics/FaQCs), FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), Trimmomatic [14] and RSeQC [73].

After the quality filtering, the clean reads are aligned to the reference genome or de novo assembly, depending on whether a reference genome is available. The former is referred as "reference-based" transcript identification. The most widely used and efficient aligners for this purpose include Bowtie [74], TopHat2 [75], STAR [19], GMAP [76], and HISAT2 [20]. If no reliable reference genome exists for the species, de novo transcript assembly can be used to identify the transcripts using tools such as Oases [77], rnaSPAdes [78], Cufflinks [79], StringTie [21, 80], SOAPdenovo-trans [81], CLC genomics workbench [82] and Trinity [83, 84]. For species with a reference genome, both mentioned strategies can be used alone or in combination.

The next step is to estimate the expression abundance of genes or transcripts based on the alignment results. Gene (RNA) expression levels can be typically expressed as Reads Per Kilobase per Million (RPKM), Fragments Per Kilobase per Million (FPKM), Reads/Counts Per Million (RPM/CPM) and Transcripts Per Million (TPM). One of approaches for quantification is to aggregate raw counts of mapped reads using tools such as HTSeq-count [85] and featureCounts [86]. The reads count tools rely on the alignment result file (SAM or BAM file) and gene feature file (GFF) as input and generate the count number of sequencing reads overlap for each of the gene features. In addition, transcriptome assemblers like Cufflinks and StringTie could be used to obtain the abundances of novel transcripts in addition to the known ones. Several other tools like RSEM [87], Salmon [88], and eXpress [89] are also commonly used to estimate transcript abundance as well as abundance of gene isoforms.

Once quantitative counts of each transcript are available, the following step is to identify differentially expressed genes (DEGs) across different samples and conditions. Multiple prevailing approaches are applied to accurately detect differentially expressed genes, which include count-based techniques like DESeq2 [90], limma [91], and edgeR [92], assembly-based techniques like Cuffdiff [93] and Ballgown [94], or sleuth [95] which performs differential

analysis on alignment-free quantifications. In addition, many genes involved in the same cellular pathways often share similar expression profile and the detection of co-expression genes across different growth conditions has received broad attention. The weighted gene co-expression network analysis (WGCNA) is one of the most used tools to identify co-expressed groups of genes [96].

In addition to the above quantitative analysis, RNA-Seq data can also be used to identify important genomic and transcriptomic variations. GATK [97] is the best practices workflow to call variants such as SNPs and indels from RNA-seq data. Furthermore, a useful tool is vcftools [98], which is a program package designed for working with VCF files and used to filter out specific variants or summarize variants etc. The widely used applications for functional annotation of variants include SnpEff [99] and ANNOVAR [100].

Transcriptome analysis is also commonly applied to identify target genes of a specific transcription factor (TF). This is often done by firstly identifying the DEGs through comparing transcriptomic profiles of a TF mutant and the corresponding reference strain. Next, the specific motif enriched in promoters of these DEGs could be used to infer the TF binding consensus sequence. Several tools are widely used for the motif analysis such as MEME [101] and RSAT [102].

The commonly used bioinformatics tools for each aspect of RNA-seq analysis are summarized in Table 3.

**Table 3. The widely used tools in transcriptomics analysis.**

| Tools | Quick links |
|---|---|
| **Quality control** | |
| fastQC | https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ |
| Trimmomatic | http://www.usadellab.org/cms/?page=trimmomatic |
| Fastx_trimmer | https://rdrr.io/github/abshah/RADseqR/man/fastx_trimmer.html |
| RSeQC | http://rseqc.sourceforge.net/ |
| **Alignment** | |
| HISAT2 | http://daehwankimlab.github.io/hisat2/ |
| StringTie | https://ccb.jhu.edu/software/stringtie/ |
| STAR | https://github.com/alexdobin/STAR |
| SOAPdenovo-trans | https://github.com/aquaskyline/SOAPdenovo-Trans |
| Cufflinks | http://cole-trapnell-lab.github.io/cufflinks/ |
| **Variant calling and genotyping** | |
| GATK | https://gatk.broadinstitute.org/hc/en-us |
| **Variant annotation** | |
| SnpEff | http://snpeff.sourceforge.net/ |
| ANNOVAR | https://doc-openbio.readthedocs.io/projects/annovar/en/latest/ |
| **Calculate reads count** | |
| HTSeq-count | https://htseq.readthedocs.io/en/release_0.11.1/count.html |
| featureCounts | http://subread.sourceforge.net |
| **Differential expression** | |
| DEseq2 | https://bioconductor.org/packages/release/bioc/html/DESeq2.html |
| edgeR | https://bioconductor.org/packages/release/bioc/html/edgeR.html |
| **Visualization** | |
| IGV | https://software.broadinstitute.org/software/igv/ |
| **Gene annotation** | |
| GO | https://go.princeton.edu/ |
| KEGG | https://www.genome.jp/kegg/ |
| GSEA | https://www.gsea-msigdb.org/gsea/index.jsp |
| BLAST2GO | https://www.blast2go.com/ |
| Pfam | https://pfam.xfam.org/ |
| WGCNA | http://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/ |

(**Table 3**. Continued)

| Tools | Quick links |
|---|---|
| **Utility tools** | |
| Bedtools | https://bedtools.readthedocs.io/en/latest/ |
| SAMtools | https://www.htslib.org/ |
| Vcftools | http://vcftools.sourceforge.net/ |
| **Online platform** | |
| Galaxy | https://usegalaxy.org/ |

### *3.1.2 Examples of fungal transcriptomics studies*

RNA-seq analysis has been widely applied in a broad range of fungal studies and greatly advanced our understanding of fungal physiology and genetic regulation. For example, RNA-seq has been widely deployed to study fungal plant biomass degradation in diverse fungal species across different growth conditions. These studies range from the dynamic transcriptional responses of fungus to lignocellulose [103-106] and lignocellulose-derived compounds [107-109], to complex transcriptional regulation network [110-112]. In addition, sophisticated bioinformatics analysis of strand-specific RNA-seq data have indicated that antisense transcripts [103, 113, 114] and RNA editing [115] may play a role during fungal grown on plant biomass. The combined analysis of large-scale transcriptome datasets across different fungal species and growth conditions have enabled us to detect conserved genes and pathways during fungal development or their adaption to natural biotopes [116, 117].

## *3.2 Proteomics*

The concept of proteome was first proposed in 1994 by Marc Wilkins [118]. The proteome is the entire set of proteins in a given cell, tissue or biological sample, at a precise developmental or cellular phase. Proteomics, the global analysis of proteins, aims for qualitative and quantitative measurements of the production, function and regulation of the entire set of proteins encoded by the genome at a specific condition. Mass spectrometry (MS) is the most

common technique for proteomics analysis and has been widely applied in fungal research [119-124]. Due to the increasing throughput and accuracy of current MS instruments, the importance of bioinformatics analysis of vast amounts of proteomics data is becoming evident. Here we mainly focus on the proteomics related bioinformatics methods, while only give a very short summary of experimental methods. The experimental methods of proteomics can be found in other well-written reviews [125-127].



The most widely-adopted proteomics methodology for large-scale proteomics analysis is the MS based bottom-up proteomics approach [128-130], while the top-down and middle-down proteomics strategy is still in the development stage [131-133]. In bottom-up proteomics, we do not use MS to analyze the intact protein, but analyze a peptide mixture digested from proteins. Then MS detecting signals are used to infer peptide information, which can be further used to infer protein information. Owing to the recent advances in instrumentation, sample preparation and computational algorithms, proteomics researchers can routinely identify and quantify thousands of proteins in a single experiment. A generic proteomics workflow can be broadly categorized into three major steps, i.e., (i.) sample preparation, (ii.) MS data acquisition and (iii.) data analysis or post-MS data acquisition. Proteins extracted from a biological sample are digested with a specific enzyme (typically trypsin). The resulting peptides are separated by liquid chromatography. Then the ionized peptides are introduced to mass spectrometry that detects the mass and ion intensity for peptides and their fragments. At the last step, via computational software, peptides and proteins are identified and quantified. The possible biological function can be further annotated via various bioinformatics analysis. Many strategies used in transcriptomic data analysis for biological function mining can be directly or slightly modified to be applied in proteomics data analysis.

*3.2.1 Bioinformatics workflow and tools for proteomics analysis*

Today, proteomics is a multi-disciplinary field, which has been extended from the initial large-scale protein identification, to quantification, and to

the characterization of complex protein-protein interactions. In parallel, corresponding bioinformatics tools are also evolving so as to address problems arising from specific branches of proteomics research. In fact, intensive developments in these areas have led to the creation of a new field named "computational proteomics" dedicated completely to this pursuit [134-136]. In this section, we mainly focus on data analysis methods that are commonly used in most fungal proteomics studies including: (i.) peptide and protein identification; (ii.) quantification of proteins; and (iii.) biological interpretation of proteomics results.

1) Peptide and protein identification

Normally, it takes two steps to infer protein identities from MS data. The first step is by matching a MS/MS spectrum to a peptide sequence, known as a peptide spectrum match (PSM). Subsequently, peptide sequences obtained from PSMs are used to infer proteins. As with any large-scale studies such as genomics and transcriptomics, statistical models play a key role for controlling the false discovery rate (FDR) for both peptide and protein identification. A number of computational algorithms and software solutions have been developed to assign peptide sequences to fragment ion spectra. These tools can be classified into three categories according reviews [137-139]:

1. Database searching: peptide sequences are identified by correlating an acquired $MS^2$ spectrum with a theoretical spectrum predicted from peptides obtained from *in silico* digestion of a protein sequence database [140, 141]. Alternatively, in the so-called spectral-library search approach a fragment ion spectrum can be correlated with a library of experimental MS/MS spectra identified in previous experiments [142-144].

2. *De novo* sequencing: peptide sequences are directly inferred from fragment ion spectra without reference to any database information. Many newly developed tools designed for *de novo* peptide sequencing have introduced advanced machine learning algorithms and showed promising results, such

as pNovo 3 [145], Novor [146], and DeepNovo [147].

3. Hybrid approaches that combine database search and *de novo* sequencing: based on the extraction of short "sequence tags" of 3-5 residues in length, followed by "error-tolerant" database searching [148-150]

A list of tools commonly used for these different types of peptide spectrum matching is provided in Table 4. Currently, database searching strategies represent the most-used method for high-throughput peptide identification.



After peptide to spectrum matching, the next step is to assemble peptide sequences into proteins. However, mapping peptide sequences to proteins is not always straightforward, as some peptide sequences are shared among more than one protein, due to sequence homology or alternative splicing of genes. To solve this problem, a common strategy is by simply reporting the parsimonious set of proteins to explain the observed peptides [174].

**Table 4. Overview of software available for peptide and protein identification.**

| Program | Website | Ref |
|---|---|---|
| **Database search** | | |
| SEQUEST | http://fields.scripps.edu/sequest/ | [141] |
| MASCOT | http://www.matrixscience.com/ | [151] |
| X!tandem | http://www.thegpm.org/tandem/ | [152] |
| Andromeda | http://www.coxdocs.org/doku.php?id=maxquant:andromeda | [153] |
| Comet | http://comet-ms.sourceforge.net/ | [154] |
| MS Amanda | https://ms.imp.ac.at/?goto=msamanda | [155] |
| MS-GF+ | http://proteomics.ucsd.edu/software-tools/ms-gf/ | [156] |
| SearchGUI | http://searchgui.googlecode.com | [157] |
| ProLuCID | http://fields.scripps.edu/yates/wp/?page_id=821 | [158] |

(**Table 4**. Continued)

| Program | Website | Ref |
|---|---|---|
| **Database search** | | |
| msFragger | https://github.com/Nesvilab/MSFragger | [159] |
| **De novo peptide sequencing** | | |
| pNovo 3 | http://pfind.ict.ac.cn/software/pNovo/index.html | [145] |
| pNovoM | http://pfind.ict.ac.cn/software/pNovoM/ | [160] |
| DeepNovo | https://github.com/nh2tran/DeepNovo | [147] |
| DeNovoGUI | http://compomics.github.io/projects/denovogui | [161] |
| Novor | https://www.rapidnovor.org/novor | [146] |
| PepNovo | http://proteomics.ucsd.edu/Software/PepNovo | [162] |
| PEAKs | https://www.bioinfor.com | [163] |
| **Hybrid approaches** | | |
| MS Dictionary | http://proteomics.ucsd.edu/Software/ | [149] |
| TagRecon | https://lab.vanderbilt.edu/msrc-bioinformatics/software/ | [150] |
| PEAKS DB | https://www.bioinfor.com/peaksdb/ | [148] |
| **Statistical validation** | | |
| PeptideProphet | http://peptideprophet.sourceforge.net/ | [164] |
| ProteinProphet | http://proteinprophet.sourceforge.net/ | [165] |
| Percolator | https://noble.gs.washington.edu/proj/percolator/ | [166] |
| Scaffold | http://www.proteomesoftware.com/ | [167] |
| **Proteomics data repositories** | | |
| Global Proteome Machine Database | http://gpmdb.thegpm.org/ | [168] |
| PeptideAtlas | http://www.peptideatlas.org/ | [169] |
| jPOST | https://jpostdb.org/ | [170] |
| ProteomeXchange Consortium | http://www.proteomexchange.org/ | [171] |
| Proteomics Identifications (PRIDE) Database | https://github.com/PRIDE-Toolsuite/pride-inspector | [172] |
| PASSEL | http://www.peptideatlas.org/passel/ | [173] |

As with other high-throughput “omics” technologies, proteomics is capable of generating massive amounts of data, albeit with a certain degree of noise, which affects peptide and protein identifications. As such, assessment of the quality of proteomics data is of utmost importance [166, 174-177]. Many probabilistic methods have been developed to provide a statistical measure

of confidence and an estimate of error rates [174]. For estimating the false-positive rate in peptide to spectrum matching, the most popular method makes use of a target-decoy analysis [178]. This method requires a decoy database, which is created by shuffling or reversing the protein sequences in the target protein database of the organism under study. It assumes that scores of spectra matched to a decoy database distribute similarly as those of incorrect matches to the original target database. The false discovery rate (FDR) of a certain score threshold can be determined by the number of PSMs in the decoy search divided by the PSM counts in the target search. Other approaches to separate the correct PSMs from incorrect ones are based on machine learning methods, such as PeptideProphet [164] and Percolator [166]. For protein inference, several statistical models have been proposed to estimate the reliability of the assigned identified protein. For instance, ProteinProphet computes a probability that a protein is present in the sample by combining the probabilities of peptides assigned to this protein [165]. In addition, several statistical methods have been specially developed to assess the confidence of the post-translation modification sites, such as Ascore [179], PTM score [180] and PhosphoRS [181].



In addition, with continually increasing of proteomics application, sharing and reusing of data become essential for the research community. Currently, a number of proteomics tools, databases and repositories have sprung up which provide a highly valuable source for bioinformatics data mining (Table 4) [182, 183].

2) Data analysis of quantitative proteomics

MS based proteomics has not only been successfully used for large-scale protein identification, but is also widely applied for quantifying global protein production [184]. Over the past decade, many quantification strategies for MS-based proteomics have been introduced [137, 174, 185, 186]. In these methods, quantitative information from mass spectra at various levels can be extracted and used for estimating protein abundance. This information can

either originate from a $MS^1$ feature (peaks in the $MS^1$ spectrum characterized by their intensity, m/z value, and the time of acquisition of the spectrum), or $MS^2$ features (fragment ion spectra) or even peptide features (the PSMs for each peptide or groups of isotopic mass peaks originating from the same peptide ion). In general, based on the particular sample preparation workflow, quantitative proteomics can be categorized in two flavors: (i.) Based on stable isotope labeling or (ii.) Based on label free methods which mainly include spectral counting (PSM-based) and determination of precursor ion signal intensities ($MS^1$-based).

Quantitative proteomics analysis is not straightforward as the intensity detected by MS may be noisy, biased or contain missing values. Instead, the next step is to apply data analysis methods to clean up the data or to correct for these biases, followed by statistically determining and confirming differentially expressed proteins [187-190]. These statistical methods address key issues in data analysis ranging from data normalization [191, 192] to imputation of missing values and also significance analysis [193]. As transcriptomics and proteomics data do bear some similarities, statistical methods successfully applied in quantitative proteomics are mostly adapted from microarray-based data analysis.

3) Biological interpretation of proteomics results

Bioinformatics plays an important role in the research of functional protein annotation, and prediction of protein-protein interactions. For function prediction of an unknown protein, one simple and common method is to compare its sequence with the known proteins by using sequence alignment tools like ClustalW [194], Jalview [195], and MAFFT [196], or using BLASTp to search the sequence against well curated databases such as NCBI [197] and Uniprot [198]. In addition, the general functional annotation of proteins, such as GO terms, KEGG, and PFAM are also useful to infer their function. Several tools can be used to achieve this goal such as Blast2GO [199], InterProScan [200] and FungiFun2 [201]. Specifically in fungi, the carbohydrate-active

enzymes (CAZymes) which are responsible for the breakdown, biosynthesis or modification of glycoconjugates, oligo- and polysaccharides, could be assigned to a particular class in the CAZy database [202] to predict their function. Furthermore, the fungal secreted protein could be predicted with several common bioinformatics tools, such as SignalP, WolfPsort [203] and Phobius [204], or can be directly retrieved from the fungal secretome databases such as FunSecKB (http://bioinformatics.ysu.edu/secretomes/fungi.php) [205] and FSD (http://fsd.snu.ac.kr/) [206]. Enrichment analysis like GO enrichment, and gene set enrichment algorithms (GSEA) can be performed to further explore the function of a selected protein list [207]. Algorithms such as *motif-x* [208] or PhosphoMotif Finder [209] analyze the sequence environment of phosphorylation modification sites in a proteomics dataset, reporting enrichment of certain amino acid motifs which can help to identify the kinases catalyzing the modification.

For protein-protein interaction, a widely used resource of interaction data is STRING [210, 211], which is not only a widely used database for protein interaction data, but also providing various other resources for literature mining. Cytoscape [212] is a convenient tool to visualize the protein-protein interaction network. With many plug-ins, Cytoscape enables further analysis of the protein network, such as the MCODE [213] for topological analysis and BiNGO for Gene Ontology enrichment analysis [214].

### *3.2.2 Examples of fungal proteomics studies*

The capability of high-throughput identification and quantification of proteins, the actual functional molecules in the cell, enables proteomics to provide complement information to genomics and transcriptomics in reflecting the cellular state. In the past decade, numerous proteomics studies were carried out alone or together with transcriptomics to investigate fungal plant biomass utilization. For instance, MS based proteomics have been widely used to monitor the fungal secretome, which revealed dynamic changes of key lignocellulolytic enzymes in response to different carbon sources [215,

216]. The comparative transcriptome and secretome analyses of the brown rot fungus *Postia placenta* and the white rot fungus *P. chrysosporium* cultured in wood containing media elucidated the different mechanisms of lignocellulose degradation mechanisms employed by these model wood decay fungi [217-220]. Comparison of secretome from different fungi showed that the similar genomic content could generate a very diverse secretome, especially the CAZyme profile, in response to plant biomass degradation [221]. In addition, several studies have indicated an important role of post-translational modification (such as phosphorylation) in regulating fungal plant biomass degradation through phospho-proteomics analysis [222, 223].

### *3.3 Metabolomics*

Metabolomics originated from metabolic profiling. Fiehn proposed "metabolomics" in 2001 and defined it as "a comprehensive and quantitative analysis of all metabolites in a biological system" [224]. Because metabolites and their concentrations directly reflect the underlying cellular biochemical activity, metabolomics is a useful and complementary approach to other omics techniques in deciphering the complexity of biological systems.

The general procedure for metabolomics analysis includes the separation, detection, identification and quantification of metabolites. Gas or liquid chromatography-mass spectrometry (GC-MS or LC-MS) and nuclear magnetic resonance (NMR) are the most frequently used tools for determining and quantifying as many metabolites as possible, including both known and unknown compounds. The analytical methodologies for metabolite profiling have been extensively reviewed in many papers and books [225-227]. The high separation efficiency, excellent reproducibility, and many well-established and available libraries for structural identification are key advantages of GC-MS analysis [228]. Typically, classes of metabolites routinely measured by GC-MS include sugars, amino acids, phosphorylated metabolites, organic acids, lipids, and amines. LC-MS is another technique commonly used in metabolomics. Sample preparation for LC-MS is simple

and it caters well for compounds that are not well suited for GC-MS, such as macromolecular metabolites, thermally labile compounds, polar compounds large, and compounds that do not volatilize easily. To date in fungi, LC-MS is a robust technique for examining and screening secondary metabolites [229]. As for NMR, it has the advantage of simplicity of sample preparation, high precision as well as high resolution. However, it suffers from relatively low sensitivity, and concentration insensitivity on detecting metabolites with large concentration differences in biological systems [230, 231], which hinders its application in fungal metabolomics [228]. Particularly for fungi, direct infusion mass spectrometry (diMS) [232] can probably be the best and most efficient approach used for the quantitative analysis of metabolites. In addition to relative easy-to-use and high resolution and accuracy, the major advantage of diMS is that each observed ion can be associated with a unique chemical formula so that the accurate mass of this ion can be determined. Therefore, diMS is applied to rapidly classify filamentous fungi. The classified species were typically identified based on their mass profile rather than a conventional phenotypic identification [233].



After the chromatographic separation of biological metabolites, a large amount of generated spectral and multivariate data needs to be analyzed by employing statistics and computational tools. Independent of the employed analytical technique, in broad terms, metabolomic profiling falls into two categories: targeted and untargeted analyses. For the targeted approach, it is restricted to defined metabolites, for example a group of particular metabolic enzymes, and can quantify and identify by monitoring a set of annotated known substances. By contrast, untargeted profiling methods try to detect all measurable compounds, including unknown or at least unidentified [234-236]. Thus, untargeted approaches have the potential of probing the entire metabolic space [237]. GC-MS could be well applied to non-targeted metabolite profiling, especially for volatile and thermally stable polar and non-polar metabolites [229].

*3.3.1 Bioinformatics workflow and tools for metabolomic analysis*

Over the last decade, several software programs for automated processing of metabolomics data have been introduced to assist data mining in metabolomics. MZmine 2 (http://mzmine.sourceforge.net/) is an open-source software for mass-spectrometry data processing, with the main focus on LC-MS data, which has been applied to both targeted and non-targeted metabolomic analyses [238]. MetAlign has been successfully used in many metabolomics studies, which can process GC-MS and LC-MS data with multiple threads for untargeted analysis [239]. OpenMS 2.0 (http://www.openms.de) provides many tools that are specifically designed for the metabolite quantification and metabolite identification in non-targeted metabolomics analysis [240]. XCMS (https://xcmsonline.scripps.edu) also processes untargeted metabolomic data, and provides a complete workflow including feature detection, retention time correction, alignment, annotation, statistical analysis, and data visualization [241]. In addition, XCMS Online is integrated with METLIN [242], a large metabolite database. Several metabolic pathways and biochemical databases are available for metabolomics analysis. Among them, BioCyc (https://biocyc.org/), MetaCyc (MetaCyc Encyclopedia of Metabolic Pathways, https://metacyc.org/), KEGG (https://www.genome.jp/kegg/) and YMDB (http://www.ymdb.ca/) are dominant databases for fungal research. However, a fungi-specific well-established metabolomics database is still lacking. Establishment of a microbial metabolomics database will accelerate the identification of compounds and species. Besides those mentioned databases, FUNGIpath (http://fungipath.i2bc.paris-saclay.fr/) appears to be a useful tool to assess fungal metabolic pathways predicted by orthology [243]. In addition, the metaP server (http://metap.helmholtz-muenchen.de/metap2/) can provide automated and standardized data analysis for quantitative metabolomics data, which is an easy-to-use and web-server-based platform covering series of analysis from data acquisition to biological interpretation [244].

Previous studies have shown that genes involved in secondary metabolite

biosynthesis are often co-regulated and physically linked into gene clusters on the chromosomes [245, 246]. Those metabolic gene clusters are referred as biosynthetic gene clusters (BGCs). BGCs in fungi encode various unexploited high-value metabolites that are closely related to food, agriculture, and medical fields. In the past few years, various bioinformatics methods have been developed for detecting gene clusters in fungi, including antiSMASH (antibiotics & Secondary Metabolite Analysis Shell) [246], BiG-SCAPE/CoRASoN [247], and SMURF (Secondary Metabolite Unknown Regions Finder) [248], MIDDAS-M (Motif-Independent *De novo* Detection Algorithm for Secondary Metabolite gene clusters) [249], CASSIS-SMIPS (Cluster Assignment by Island of Sites and Secondary Metabolites by InterProScan) [250], C-Hunter [251] and NRPSpredictor2 [252]. Genome data and transcriptome data can also be integrated in the identification of gene clusters. For instance, a comparative genomic method was applied to analyze five *Aspergillus* species and ten other filamentous fungal species, which successfully identified secondary metabolite biosynthesis gene clusters by searching for a similar order of genes and their presence in non-syntenic blocks based on synteny analysis [253]. A "multi-omics"-based method was performed on *Aspergillus nidulans*, which integrated metabolite profiling analysis and transcriptomics analysis to accurately determine fungal the secondary metabolites (SMs) gene cluster members [254]. In addition, FunGeneClusterS software [255] was developed for accurate prediction of gene clusters from genome and transcriptome data, which can predict co-regulated clusters not limited to secondary metabolism. In response to an increasing interest in the research of secondary metabolites, several databases have been developed in recent year, including ClusterMine360 (http://www.clustermine360.ca/) [256], IMG-ABC (https://img.jgi.doe.gov/cgi-bin/abc/main.cgi) [257] and MIBiG (Minimum Information about a Biosynthetic Gene Cluster, https://mibig.secondarymetabolites.org/) [258]. They provide useful information on gene clusters and related function annotation, which enable the exploitation of fungal metabolome.

*3.3.2 Examples of fungal metabolomics studies*

Metabolomics has been widely used to investigate changes of fungal metabolites profile in response to various environmental stimuli or genetic modification [259-262]. For instance metabolomics were extensively used in studies of fungal pathogen-plant interactions to understand plant defense mechanisms, such as identifying fungi, determining infection mechanisms, and detecting interactions with the host [228]. Up to now, metabolomics analyses of fungal pathogen-plant interactions are limited to several reference fungi and their host plants, such as *F. graminearum* [263, 264], *Magnaporthe oryzae* [265], *U. maydis* [266], *Rhizoctonia solani* [267, 268], *Botrytis cinerea* [269-271], *Sclerotinia sclerotiorum* [272], *Aspergillus oryzae* [273], *Penicillium digitatum* [274] and their hosts [228]. Metabolomics has also been commonly used to discover metabolic molecules related to fungi-fungi interaction. In addition, metabolomics has been increasingly used in help understanding of fungal plant biomass degradation [275, 276].

## 4. Integrative analysis of multiple omics

The single-level omics approaches (e.g. genomics, transcriptomics, proteomics, and metabolomics) have resulted in vast amounts of data and advanced our understanding of fungal biological processes. Combining multiple levels of omics information is an emerging approach, which helps uncover the hidden complex biological mechanism that may become evident only through integrative analysis of multiple molecular profiles [277]. The increase in different omics data available in public databases and newly developed bioinformatics tools accelerate the applications of integrative analysis. One of the straightforward applications of integrated multiple omics data analysis is using transcriptomics and proteomics for improving genome annotation and validating gene prediction. Other applications include directly correlating or sophisticatedly integrating different molecular profiles to discover new biological mechanisms. For instance, a comparative analysis of multiple transcriptome datasets from 10 different basidiomycetes

available in public databases showed that a large set of conserved genes, especially the CAZymes encoding genes, are highly expressed in plant biomass related substrates, suggesting that these genes are critical for fungal plant biomass conversion [117]. A large-scale co-expression analysis of *A. niger* transcriptome data was applied to improve the genome annotation and identify new function of several genes [278]. Several recent combined analyses of transcriptome, secretome and even metabolome data have showed great potential for a comprehensive understanding of the process of fungal plant biomass degradation [276, 279, 280]. In addition, more ambitious efforts have been performed to integrate different omics data, genetics and biochemical characterizations to reconstruct the networks of fungal plant biomass utilization, such as performed in *N. crassa* [281] and *A. niger* [282].



To date, several tools have been developed to address the integration of different omics data based on biochemical pathway, ontology, network and correlation (see recent review of [283]). For more advanced users with expertise in programming, a variety of tools are available for perform integrative analysis of multiple omics data, such as IntegrOmics [284], SteinerNet [285], Omics Integrator [286], MixOmics [287], SNF [288], iCluster [289], omicade4 package [290], CIMLR [291], and ShinyOmics [292]. As most of these tools are originally designed for analyzing human and plant related data, some modifications are needed to enable them to be applied in fungal studies. Recently, Miyauchi et al. designed a bioinformatics pipeline for integrative analysis of transcriptomic and secretomic data based on self-organizing map algorithm and enrichment analysis. In the test case, it successfully identified condition-specific expressed genes and secreted enzymes that are synergistically involved in fungal plant biomass degradation [293].

## 5. Challenges and perspectives for future bioinformatics analysis

The progress of genomic and post-genomic technologies during the past

decades has exponentially increased a variety of omics data and significantly enriched the list of potential genes and pathways related to fungal evolution, physiology and industrial applications. However, how to make best use of those large omics datasets remains a challenge. Firstly, the number of well-annotated genomes is severely lagging behind the total number of available draft genome sequences. Due to the remarkable genetic diversity that exists in the fungal kingdom, most gene function predictions based on sequence homology to well-annotated reference fungi cannot guarantee an accurate gene function assignment for many of the newly sequenced genomes. A large part of the genes in most fungal genomes are simply assigned as unknown function based on computational prediction due to their low similarity with characterized genes. Since genome annotation is the starting point for analysis of genome content, the lack of well-annotated genomes could severely hinder proper interpretation of functional genomics data. For example, the accuracy of most functional enrichment analyses relies heavily on genome annotation quality. To address this issue, a gold-standard genome program covering a wide range of fungi species is ongoing [294, 295]. Transferring the annotation of gold-standard genomes to other closely related species could help to explore genome and gene function in a broader set of species, and ensure complete and consistent annotation sets. A modified gene annotation method, named RATT, can be applied using well-annotated genome and gene models as a reference to quickly transfer genome annotation [296]. Secondly, the integrative analysis of multiple omics datasets is still at an early stage. Most current integrative analyses rely on manually collecting comparable datasets from different public databases or self-generated large-omics datasets, which is time consuming. Building a central database to connect all related biological data (including omics data, phenotypes, genetics and biochemical characterizations), could help to recycle existing datasets and discover new biological mechanisms via integrate analysis of multiple datasets. In addition, the tools and algorithms for integrative analysis of fungal omics lag behind, compared to the numerous tools designed for handling single omics data, and

multiple integrative analysis tools developed for handling animal and plant related data. Thirdly, most current analyses focus only on certain sets of genes depending on the specific research interest, but future bioinformatic studies need to conduct more network analysis, which could help to identify new genes associated with specific phenotypes. For instance, co-expression [278] has been used to link unknown genes with known ones based on their similar expression patterns across a large set of transcriptomic data.



With many applications of various omics technologies and rapid expansion of omics data, bioinformatics has become an indispensable unit in today's biological research. In this chapter, we only selectively introduced a number of commonly used bioinformatic tools, which only represent a small part of total available tools. The Galaxy (https://usegalaxy.org) project, a web-based scientific analysis platform has currently collected more than 5500 tools that have been frequently used by tens of thousands of scientists across the world to analyze various omics data. In the future, we expect more advanced and user-friendly bioinformatics tools to be developed to deal with the increasing large, complex and high-dimensional biological data and to help identify new biological mechanisms.

## Acknowledgements

JL was supported by the China Scholarship Council (CSC 201909110079).

## References

1. Goffeau, A., et al., *Life with 6000 genes.* **Science**, 1996. 274(5287): p. 546-567.
2. Wood, V., et al., *The genome sequence of Schizosaccharomyces pombe.* **Nature**, 2002. 415(6874): p. 871-880.
3. Galagan, J.E., et al., *The genome sequence of the filamentous fungus Neurospora crassa.* **Nature**, 2003. 422(6934): p. 859-868.
4. Galagan, J.E., et al., *Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae.* **Nature**, 2005. 438(7071): p. 1105-1115.
5. Pel, H.J., et al., *Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88.* **Nature Biotechnology**, 2007. 25(2): p. 221-231.
6. Martinez, D., et al., *Genome sequencing and analysis of the biomass-degrading*

*fungus Trichoderma reesei (syn. Hypocrea jecorina).* **Nature Biotechnology**, 2008. 26(5): p. 553-560.

7. Dean, R.A., et al., *The genome sequence of the rice blast fungus Magnaporthe grisea.* **Nature**, 2005. 434(7036): p. 980-986.
8. Kamper, J., et al., *Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis.* **Nature**, 2006. 444(7115): p. 97-101.
9. Cuomo, C.A., et al., *The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization.* **Science**, 2007. 317(5843): p. 1400-1402.
10. Martinez, D., et al., *Genome sequence of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78.* **Nature Biotechnology**, 2004. 22(6): p. 695-700.
11. Sharma, K.K., *Fungal genome sequencing: basic biology to biotechnology.* **Critical Reviews in Biotechnology**, 2016. **36**(4): p. 743-759.
12. Morin, E., et al., *Genome sequence of the button mushroom Agaricus bisporus reveals mechanisms governing adaptation to a humic-rich ecological niche.* **Proceedings of the National Academy of Sciences**, 2012. 109(43): p. 17501-17506.
13. Besser, J., et al., *Next-generation sequencing technologies and their application to the study and control of bacterial infections.* **Clinical Microbiology and Infection**, 2018. 24(4): p. 335-341.
14. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* **Bioinformatics**, 2014. 30(15): p. 2114-2120.
15. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* **EMBnet. journal**, 2011. 17(1): p. 10-12.
16. Patel, R.K. and M. Jain, *NGS QC Toolkit: a toolkit for quality control of next generation sequencing data.* **PLoS One**, 2012. 7(2): p. e30619.
17. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* **Nature Methods**, 2012. 9(4): p. 357-359.
18. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* **Bioinformatics**, 2010. 26(5): p. 589-595.
19. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* **Bioinformatics**, 2013. **29**(1): p. 15-21.
20. Kim, D., B. Langmead, and S.L. Salzberg, *HISAT: a fast spliced aligner with low memory requirements.* **Nature Methods**, 2015. 12(4): p. 357-360.
21. Pertea, M., et al., *StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.* **Nature Biotechnology**, 2015. 33(3): p. 290-295.
22. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* **Briefings in Bioinformatics**, 2013. 14(2): p. 178-192.

23. Giardine, B., et al., *Galaxy: A platform for interactive large-scale genome analysis.* **Genome Research**, 2005. 15(10): p. 1451-1455.
24. Pop, M., *Genome assembly reborn: recent computational challenges.* **Briefings in Bioinformatics**, 2009. 10(4): p. 354-366.
25. Simpson, J.T., et al., *ABySS: a parallel assembler for short read sequence data.* **Genome Research**, 2009. 19(6): p. 1117-1123.
26. Li, R., et al., *De novo assembly of human genomes with massively parallel short read sequencing.* **Genome Research**, 2010. 20(2): p. 265-272.
27. Luo, R., et al., *SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.* **GigaScience**, 2012. 1(1): p. 18.
28. Gnerre, S., et al., *High-quality draft assemblies of mammalian genomes from massively parallel sequence data.* **Proceedings of the National Academy of Sciences**, 2011. 108(4): p. 1513-1518.
29. Koren, S., et al., *Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.* **Genome Research**, 2017. 27(5): p. 722-736.
30. Chin, C.S., et al., *Phased diploid genome assembly with single-molecule real-time sequencing.* **Nature Methods**, 2016. 13(12): p. 1050-1054.
31. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* **Genome Research**, 2008. 18(5): p. 821-829.
32. Salzberg, S.L., et al., *GAGE: A critical evaluation of genome assemblies and assembly algorithms.* **Genome Research**, 2012. 22(3): p. 557-567.
33. Gurevich, A., et al., *QUAST: quality assessment tool for genome assemblies.* **Bioinformatics**, 2013. 29(8): p. 1072-1075.
34. Simao, F.A., et al., *BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.* **Bioinformatics**, 2015. 31(19): p. 3210-3212.
35. Yang, L.A., et al., *SQUAT: a Sequencing Quality Assessment Tool for data quality assessments of genome assemblies.* **BMC Genomics**, 2019. 19(Suppl 9): p. 238.
36. Manchanda, N., et al., *GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations.* **BMC Genomics**, 2020. 21(1): p. 1-9.
37. Burge, C. and S. Karlin, *Prediction of complete gene structures in human genomic DNA.* **Journal of Molecular Biology**, 1997. 268(1): p. 78-94.
38. Lukashin, A.V. and M. Borodovsky, *GeneMark.hmm: new solutions for gene finding.* **Nucleic Acids Research**, 1998. 26(4): p. 1107-1115.
39. Korf, I., *Gene finding in novel genomes.* **BMC Bioinformatics**, 2004. 5: p. 59.
40. Majoros, W.H., M. Pertea, and S.L. Salzberg, *TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.* **Bioinformatics**, 2004. 20(16): p. 2878-2879.
41. Cantarel, B.L., et al., *MAKER: An easy-to-use annotation pipeline designed for*

*emerging model organism genomes.* **Genome Research**, 2008. 18(1): p. 188-196.

42. Stanke, M. and S. Waack, *Gene prediction with a hidden Markov model and a new intron submodel.* **Bioinformatics**, 2003. 19: p. Ii215-Ii225.
43. Birney, E., M. Clamp, and R. Durbin, *GeneWise and Genomewise.* **Genome Research**, 2004. 14(5): p. 988-995.
44. Park, J., et al., *CFGP: a web-based, comparative fungal genomics platform.* **Nucleic Acids Research**, 2008. 36(Database issue): p. D562-D571.
45. Griffiths-Jones, S., et al., *Rfam: an RNA family database.* **Nucleic Acids Research**, 2003. 31(1): p. 439-441.
46. Griffiths-Jones, S., et al., *Rfam: annotating non-coding RNAs in complete genomes.* **Nucleic Acids Research**, 2005. 33: p. D121-D124.
47. Eddy, S.R., *A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure.* **BMC Bioinformatics**, 2002. 3: p. 18.
48. Liu, C.N., et al., *NONCODE: an integrated knowledge database of non-coding RNAs.* **Nucleic Acids Research**, 2005. 33: p. D112-D115.
49. Amaral, P.P., et al., *lncRNAdb: a reference database for long noncoding RNAs.* **Nucleic Acids Research**, 2011. 39: p. D146-D151.
50. Li, L., C.J. Stoeckert, Jr., and D.S. Roos, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* **Genome Research**, 2003. 13(9): p. 2178-2189.
51. Emms, D.M. and S. Kelly, *OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy.* **Genome Biology**, 2015. 16: p. 157.
52. Cantarel, B.L., et al., *The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics.* **Nucleic Acids Research**, 2009. 37(Database issue): p. D233-D238.
53. Zhang, H., et al., *dbCAN2: a meta server for automated carbohydrate-active enzyme annotation.* **Nucleic Acids Research**, 2018. 46(W1): p. W95-W101.
54. Sheppard, T.K., et al., *The Saccharomyces Genome Database Variant Viewer.* **Nucleic Acids Research**, 2016. 44(D1): p. D698-D702.
55. Skrzypek, M.S., et al., *The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data.* **Nucleic Acids Research**, 2017. 45(D1): p. D592-D596.
56. Cerqueira, G.C., et al., *The Aspergillus Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations.* **Nucleic Acids Research**, 2014. **42**(D1): p. D705-D710.
57. Amselem, J., et al., *Genomic analysis of the necrotrophic fungal pathogens Sclerotinia sclerotiorum and Botrytis cinerea.* **PLOS Genetics**, 2011. **7**(8): p. e1002230.

58. de Wit, P.J., et al., *The genomes of the fungal plant pathogens Cladosporium fulvum and Dothistroma septosporum reveal adaptation to different hosts and lifestyles but also signatures of common ancestry.* **PLoS Genetics**, 2012. 8(11): p. e1003088.
59. Ohm, R.A., et al., *Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi.* **PLOS Pathogens**, 2012. 8(12): p. e1003037.
60. Floudas, D., et al., *The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes.* **Science**, 2012. 336(6089): p. 1715-1719.
61. Peng, M., et al., *The draft genome sequence of the ascomycete fungus Penicillium subrubescens reveals a highly enriched content of plant biomass related CAZymes compared to related fungi.* **Journal of Biotechnology,** 2017. 246: p. 1-3.
62. Rytioja, J., et al., *Plant-polysaccharide-degrading enzymes from basidiomycetes.* **Microbiology and Molecular Biology Reviews**, 2014. 78(4): p. 614-649.
63. Berka, R.M., et al., *Comparative genomic analysis of the thermophilic biomass-degrading fungi Myceliophthora thermophila and Thielavia terrestris.* **Nature Biotechnology**, 2011. 29(10): p. 922-927.
64. Acton, E., et al., *Comparative functional genomic screens of three yeast deletion collections reveal unexpected effects of genotype in response to diverse stress.* **Open Biology**, 2017. 7(6): p. 160330.
65. de Vries, R.P., et al., *Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus Aspergillus.* **Genome Biology**, 2017. 18(1): p. 28.
66. Kjærbølling, I., et al., *A comparative genomics study of 23 Aspergillus species from section Flavi.* **Nature Communications**, 2020. 11(1): p. 1106.
67. Vesth, T.C., et al., *Investigation of inter-and intraspecies variation through genome sequencing of Aspergillus section Nigri.* **Nature Genetics**, 2018. 50(12): p. 1688-1695.
68. Kubicek, C.P., et al., *Evolution and comparative genomics of the most common Trichoderma species.* **BMC Genomics**, 2019. 20(1): p. 485.
69. Riley, R., et al., *Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi.* **Proceedings of the National Academy of Sciences**, 2014. 111(27): p. 9923-9928.
70. Stajich, J.E., *Fungal genomes and insights into the evolution of the kingdom.* **The Fungal Kingdom**, 2017: p. 619-633.
71. Chu, Y. and D.R. Corey, *RNA sequencing: platform selection, experimental design, and data interpretation.* **Nucleic Acid Therapeutics**, 2012. 22(4): p. 271-274.
72. Martin, J.A. and Z. Wang, *Next-generation transcriptome assembly.* **Nature Reviews Genetics**, 2011. 12(10): p. 671-682.
73. Wang, L., S. Wang, and W. Li, *RSeQC: quality control of RNA-seq experiments.*

**Bioinformatics**, 2012. 28(16): p. 2184-2185.

74. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* **Genome Biology**, 2009. 10(3): p. R25.
75. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* **Genome Biology**, 2013. 14(4): p. R36.
76. Wu, T.D. and C.K. Watanabe, *GMAP: a genomic mapping and alignment program for mRNA and EST sequences.* **Bioinformatics**, 2005. 21(9): p. 1859-1875.
77. Schulz, M.H., et al., *Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels.* **Bioinformatics**, 2012. 28(8): p. 1086-1092.
78. Bushmanova, E., et al., *rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data.* **GigaScience**, 2019. 8(9): p. giz100.
79. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* **Nature Biotechnology**, 2010. 28(5): p. 511.
80. Kovaka, S., et al., *Transcriptome assembly from long-read RNA-seq alignments with StringTie2.* **Genome Biology**, 2019. 20(1): p. 278.
81. Xie, Y., et al., *SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads.* **Bioinformatics**, 2014. 30(12): p. 1660-1666.
82. Kumar, S. and M.L. Blaxter, *Comparing de novo assemblers for 454 transcriptome data.* **BMC Genomics**, 2010. 11: p. 571.
83. Grabherr, M.G., et al., *Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data.* **Nature Biotechnology**, 2011. 29(7): p. 644.
84. Haas, B.J., et al., *De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.* **Nature Protocols**, 2013. 8(8): p. 1494.
85. Anders, S., P.T. Pyl, and W. Huber, *HTSeq-a Python framework to work with high-throughput sequencing data.* **Bioinformatics**, 2015. 31(2): p. 166-169.
86. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* **Bioinformatics**, 2014. 30(7): p. 923-930.
87. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* **BMC Bioinformatics**, 2011. 12(1): p. 323.
88. Patro, R., et al., *Salmon provides fast and bias-aware quantification of transcript expression.* **Nature Methods**, 2017. 14(4): p. 417-419.
89. Roberts, A. and L. Pachter, *Streaming fragment assignment for real-time analysis of sequencing experiments.* **Nature Methods**, 2013. 10(1): p. 71-73.
90. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* **Genome Biology**, 2014. 15(12): p. 550.

91. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* **Nucleic Acids Research**, 2015. 43(7): p. e47.
92. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* **Bioinformatics**, 2010. 26(1): p. 139-140.
93. Trapnell, C., et al., *Differential analysis of gene regulation at transcript resolution with RNA-seq.* **Nature Biotechnology**, 2013. 31(1): p. 46.
94. Frazee, A.C., et al., *Ballgown bridges the gap between transcriptome assembly and expression analysis.* **Nature Biotechnology**, 2015. 33(3): p. 243-246.
95. Pimentel, H., et al., *Differential analysis of RNA-seq incorporating quantification uncertainty.* **Nature Methods**, 2017. 14(7): p. 687-690.
96. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* **BMC Bioinformatics**, 2008. 9: p. 559.
97. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* **Nature Genetics**, 2011. 43(5): p. 491-498.
98. Danecek, P., et al., *The variant call format and VCFtools.* **Bioinformatics**, 2011. 27(15): p. 2156-2158.
99. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* **Fly**, 2012. 6(2): p. 80-92.
100. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* **Nucleic Acids Research**, 2010. 38(16): p. e164-e164.
101. Bailey, T.L., et al., *The MEME Suite.* **Nucleic Acids Research**, 2015. 43(W1): p. W39-W49.
102. Medina-Rivera, A., et al., *RSAT 2015: Regulatory Sequence Analysis Tools.* **Nucleic Acids Research**, 2015. 43(W1): p. W50-W56.
103. Delmas, S., et al., *Uncovering the genome-wide transcriptional responses of the filamentous fungus Aspergillus niger to lignocellulose using RNA sequencing.* **PLoS Genetics**, 2012. 8(8): p. e1002875.
104. Ries, L., et al., *Genome-wide transcriptional response of Trichoderma reesei to lignocellulose using RNA sequencing and comparison with Aspergillus niger.* **BMC Genomics**, 2013. 14(1): p. 541.
105. Daly, P., et al., *Transcriptomic responses of mixed cultures of ascomycete fungi to lignocellulose using dual RNA-seq reveal inter-species antagonism and limited beneficial effects on CAZyme expression.* **Fungal Genetics and Biology**, 2017. 102: p. 4-21.
106. Borin, G.P., et al., *Comparative transcriptome analysis reveals different strategies for degradation of steam-exploded sugarcane bagasse by Aspergillus niger and*

*Trichoderma reesei.* **BMC Genomics**, 2017. 18(1): p. 501.

107. Casado López, S., et al., *Induction of genes encoding plant cell wall-degrading carbohydrate-active enzymes by lignocellulose-derived monosaccharides and cellobiose in the white-rot fungus Dichomitus squalens.* **Applied and Environmental Microbiology**, 2018. 84(11): p. e00403- e00418.
108. Lubbers, R.J., et al., *Discovery of novel p-hydroxybenzoate-m-hydroxylase, protocatechuate 3, 4 ring-cleavage dioxygenase, and hydroxyquinol 1, 2 ring-cleavage dioxygenase from the filamentous fungus Aspergillus niger.* **ACS Sustainable Chemistry & Engineering**, 2019. 7(23): p. 19081-19089.
109. Kowalczyk, J.E., et al., *The white-rot basidiomycete Dichomitus squalens shows highly specific transcriptional response to lignocellulose-related aromatic compounds.* **Frontiers in Bioengineering and Biotechnology**, 2019. 7: p. 229.
110. Kowalczyk, J.E., et al., *Combinatorial control of gene expression in Aspergillus niger grown on sugar beet pectin.* **Scientific Reports**, 2017. 7(1): p. 12356.
111. Li, Z., et al., *Synergistic and dose-controlled regulation of cellulase gene expression in Penicillium oxalicum.* **PLOS Genetics**, 2015. 11(9): p. e1005509.
112. Craig, J.P., et al., *Direct target network of the Neurospora crassa plant cell wall deconstruction regulators CLR-1, CLR-2, and XLR-1.* **mBio**, 2015. 6(5): p. e01452-15.
113. Sibthorp, C., et al., *Transcriptome analysis of the filamentous fungus Aspergillus nidulans directed to the global identification of promoters.* **BMC Genomics**, 2013. 14: p. 847.
114. Solomon, K.V., et al., *Catabolic repression in early-diverging anaerobic fungi is partially mediated by natural antisense transcripts.* **Fungal Genetics and Biology**, 2018. 121: p. 1-9.
115. Wu, B., et al., *Substrate-specific differential gene expression and RNA editing in the brown rot fungus Fomitopsis pinicola.* **Applied and Environmental Microbiology**, 2018. 84(16): p. e00991-18.
116. Krizsan, K., et al., *Transcriptomic atlas of mushroom development reveals conserved genes behind complex multicellularity in fungi.* **Proceedings of the National Academy of Sciences**, 2019. 116(15): p. 7409-7418.
117. Peng, M., et al., *Comparative analysis of basidiomycete transcriptomes reveals a core set of expressed genes encoding plant biomass degrading enzymes.* **Fungal Genetics and Biology**, 2018. 112: p. 40-46.
118. Wasinger, V.C., et al., *Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium.* **Electrophoresis**, 1995. 16(1): p. 1090-1094.
119. Chalupova, J., et al., *Identification of fungal microorganisms by MALDI-TOF mass spectrometry.* **Biotechnology Advances**, 2014. **32**(1): p. 230-241.
120. Patel, R., *A moldy application of MALDI: MALDI-ToF mass spectrometry for fungal*

*identification.* **Journal of Fungi**, 2019. **5**(1): p. 4.

121. Studer, R.A., et al., *Evolution of protein phosphorylation across 18 fungal species.* **Science**, 2016. 354(6309): p. 229-232.
122. Ball, B., et al., *Mass spectrometry-based proteomics of fungal pathogenesis, host–fungal interactions, and antifungal development.* **Journal of Fungi**, 2019. 5(2): p. 52.
123. Kumar, M., et al., *Genomic and proteomic analysis of lignin degrading and polyhydroxyalkanoate accumulating β-proteobacterium Pandoraea sp. ISTKB.* **Biotechnology for Biofuels**, 2018. 11(1): p. 154.
124. Mäkelä, M.R., et al., *Genomic and exoproteomic diversity in plant biomass degradation approaches among Aspergilli.* **Studies in Mycology**, 2018. 91: p. 79-99.
125. Aslam, B., et al., *Proteomics: technologies and their applications.* **Journal of Chromatographic Science**, 2017. 55(2): p. 182-196.
126. Vidova, V. and Z. Spacil, *A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition.* **Analytica Chimica Acta**, 2017. 964: p. 7-23.
127. de Oliveira, J.M. and L.H. de Graaff, *Proteomics of industrial fungi: trends and insights for biotechnology.* **Applied Microbiology and Biotechnology**, 2011. 89(2): p. 225-237.
128. Mayne, J., et al., *Bottom-up proteomics (2013–2015): keeping up in the era of systems biology.* **Analytical Chemistry**, 2016. 88(1): p. 95-121.
129. Gillet, L.C., A. Leitner, and R. Aebersold, *Mass spectrometry applied to bottom-up proteomics: entering the high-throughput era for hypothesis testing.* **Annual Review of Analytical Chemistry** (Palo Alto Calif), 2016. 9(1): p. 449-472.
130. Zhang, Y., et al., *Protein analysis by shotgun/bottom-up proteomics.* **Chemical Reviews**, 2013. 113(4): p. 2343-2394.
131. Catherman, A.D., O.S. Skinner, and N.L. Kelleher, *Top down proteomics: facts and perspectives.* **Biochemical and Biophysical Research Communications**, 2014. 445(4): p. 683-693.
132. Donnelly, D.P., et al., *Best practices and benchmarks for intact protein analysis for top-down mass spectrometry.* **Nature Methods**, 2019. 16(7): p. 587-594.
133. Toby, T.K., L. Fornelli, and N.L. Kelleher, *Progress in top-down proteomics and the analysis of proteoforms.* **Annual Review of Analytical Chemistry** (Palo Alto Calif), 2016. **9**(1): p. 499-519.
134. Kopczynski, D., A. Sickmann, and R. Ahrends, *Computational proteomics tools for identification and quality control.* **Journal of Biotechnology**, 2017. 261: p. 126-130.
135. van den Toorn, H., *Computational proteomics: from numbers to biology*. 2017,

**Utrecht University**.

136. Sarkar, D. and S. Saha, *Computational Proteomics*, in *Systems Biology Application in Synthetic Biology*. 2016, **Springer**, New Delhi. p. 11-20.
137. Chen, C., et al., *Bioinformatics methods for mass spectrometry-based proteomics data analysis.* **International Journal of Molecular Sciences**, 2020. 21(8): p. 2873.
138. Hoopmann, M.R. and R.L. Moritz, *Current algorithmic solutions for peptide-based proteomics data generation and identification.* **Current Opinion in Biotechnology**, 2013. 24(1): p. 31-38.
139. Kertész-Farkas, A., et al., *Database searching in mass spectrometry based proteomics.* **Current Bioinformatics**, 2012. 7(2): p. 221-230.
140. MacCoss, M.J., et al., *Shotgun identification of protein modifications from protein complexes and lens tissue.* **Proceedings of the National Academy of Sciences**, 2002. 99(12): p. 7900-7905.
141. Eng, J.K., A.L. McCormack, and J.R. Yates, *An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.* **Journal of the American Society for Mass Spectrometry**, 1994. 5(11): p. 976-989.
142. Dasari, S., et al., *Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment.* **Journal of Proteome Research**, 2012. 11(3): p. 1686-1695.
143. Tharakan, R., N. Edwards, and D.R. Graham, *Data maximization by multipass analysis of protein mass spectra.* **Proteomics**, 2010. 10(6): p. 1160-1171.
144. Ning, K., D. Fermin, and A.I. Nesvizhskii, *Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets.* **Proteomics**, 2010. 10(14): p. 2712-2718.
145. Yang, H., et al., *pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework.* **Bioinformatics**, 2019. 35(14): p. i183-i190.
146. Ma, B., *Novor: real-time peptide de novo sequencing software.* **Journal of the American Society for Mass Spectrometry**, 2015. 26(11): p. 1885-1894.
147. Tran, N.H., et al., *De novo peptide sequencing by deep learning.* **Proceedings of the National Academy of Sciences**, 2017. 114(31): p. 8247-8252.
148. Zhang, J., et al., *PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification.* **Molecular & Cellular Proteomics**, 2012. 11(4): p. M111 010587.
149. Kim, S., et al., *Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra.* **Molecular & Cellular Proteomics**, 2009. 8(1): p. 53-69.
150. Dasari, S., et al., *TagRecon: high-throughput mutation identification through sequence tagging.* **Journal of Proteome Research**, 2010. 9(4): p. 1716-1726.
151. Perkins, D.N., et al., *Probability-based protein identification by searching*

*sequence databases using mass spectrometry data.* **ELECTROPHORESIS: An International Journal**, 1999. 20(18): p. 3551-3567.

152. Craig, R. and R.C. Beavis, *A method for reducing the time required to match protein sequences with tandem mass spectra.* **Rapid Communications in Mass Spectrometry**, 2003. 17(20): p. 2310-2316.
153. Cox, J., et al., *Andromeda: a peptide search engine integrated into the MaxQuant environment.* **Journal of Proteome Research**, 2011. 10(4): p. 1794-1805.
154. Eng, J.K., T.A. Jahan, and M.R. Hoopmann, *Comet: an open-source MS/MS sequence database search tool.* **Proteomics**, 2013. 13(1): p. 22-24.
155. Dorfer, V., et al., *MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra.* **Journal of Proteome Research**, 2014. 13(8): p. 3679-3684.
156. Kim, S. and P.A. Pevzner, *MS-GF+ makes progress towards a universal database search tool for proteomics.* **Nature Communications**, 2014. 5: p. 5277.
157. Vaudel, M., et al., *SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X! Tandem searches.* **Proteomics**, 2011. 11(5): p. 996-999.
158. Xu, T., et al., *ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity.* **Journal of Proteomics**, 2015. 129: p. 16-24.
159. Kong, A.T., et al., *MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics.* **Nature Methods**, 2017. 14(5): p. 513.
160. Yang, H., et al., *Precision de novo peptide sequencing using mirror proteases of Ac-LysargiNase and trypsin for large-scale proteomics.* **Molecular & Cellular Proteomics**, 2019. 18(4): p. 773-785.
161. Muth, T., et al., *DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra.* **Journal of Proteome Research**, 2014. 13(2): p. 1143-1146.
162. Frank, A. and P. Pevzner, *PepNovo: de novo peptide sequencing via probabilistic network modeling.* **Analytical Chemistry**, 2005. 77(4): p. 964-973.
163. Ma, B., et al., *PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry.* **Rapid Communications in Mass Spectrometry**, 2003. 17(20): p. 2337-2342.
164. Keller, A., et al., *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.* **Analytical Chemistry**, 2002. 74(20): p. 5383-5392.
165. Nesvizhskii, A.I., et al., *A statistical model for identifying proteins by tandem mass spectrometry.* **Analytical Chemistry**, 2003. 75(17): p. 4646-4658.
166. Kall, L., et al., *Semi-supervised learning for peptide identification from shotgun proteomics datasets.* **Nature Methods**, 2007. 4(11): p. 923-925.

167. Searle, B.C., *Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies.* **Proteomics**, 2010. 10(6): p. 1265-1269.
168. Craig, R., J.P. Cortens, and R.C. Beavis, *Open source system for analyzing, validating, and storing protein identification data.* **Journal of Proteome Research**, 2004. 3(6): p. 1234-1242.
169. Desiere, F., et al., *Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry.* **Genome Biology**, 2005. 6(1).
170. Moriya, Y., et al., *The jPOST environment: an integrated proteomics data repository and database.* **Nucleic Acids Research**, 2019. 47(D1): p. D1218-D1224.
171. Vizcaino, J.A., et al., *ProteomeXchange provides globally coordinated proteomics data submission and dissemination.* **Nature Biotechnology**, 2014. 32(3): p. 223-226.
172. Perez-Riverol, Y., et al., *PRIDE inspector toolsuite: moving toward a universal visualization tool for proteomics data standard formats and quality assessment of ProteomeXchange datasets.* **Molecular & Cellular Proteomics**, 2016. 15(1): p. 305-317.
173. Farrah, T., et al., *PASSEL: The P eptide A tlas SRM experiment library.* **Proteomics**, 2012. 12(8): p. 1170-1175.
174. Nesvizhskii, A.I., O. Vitek, and R. Aebersold, *Analysis and validation of proteomic data generated by tandem mass spectrometry.* **Nature Methods**, 2007. 4(10): p. 787-797.
175. Gauci, S., et al., *Lys-N and trypsin cover complementary parts of the phosphoproteome in a refined SCX-based approach.* **Analytical Chemistry**, 2009. 81(11): p. 4493-4501.
176. Peng, M., et al., *Protease bias in absolute protein quantitation.* **Nature Methods**, 2012. 9(6): p. 524-525.
177. Walzer, M., et al., *qcML: an exchange format for quality control metrics from mass spectrometry experiments.* **Molecular & Cellular Proteomics**, 2014. 13(8): p. 1905-1913.
178. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* **Nature Methods**, 2007. 4(3): p. 207-214.
179. Beausoleil, S.A., et al., *A probability-based approach for high-throughput protein phosphorylation analysis and site localization.* **Nature Biotechnology**, 2006. 24(10): p. 1285-1292.
180. Olsen, J.V., et al., *Global, in vivo, and site-specific phosphorylation dynamics in signaling networks.* **Cell**, 2006. 127(3): p. 635-648.
181. Taus, T., et al., *Universal and confident phosphorylation site localization using phosphoRS.* **Journal of Proteome Research**, 2011. 10(12): p. 5354-5362.

182. Martens, L. and J.A. Vizcaino, *A golden age for working with public proteomics data.* **Trends in Biochemical Sciences**, 2017. 42(5): p. 333-341.
183. Perez-Riverol, Y., et al., *Making proteomics data accessible and reusable: current state of proteomics databases and repositories.* **Proteomics**, 2015. 15(5-6): p. 930-949.
184. Ong, S.E. and M. Mann, *Mass spectrometry-based proteomics turns quantitative.* **Nature Chemical Biology**, 2005. 1(5): p. 252-262.
185. Mirza, S.P., *Quantitative mass spectrometry-based approaches in cardiovascular research.* **Circulation: Cardiovascular Genetics**, 2012. 5(4): p. 477.
186. Lindemann, C., et al., *Strategies in relative and absolute quantitative mass spectrometry based proteomics.* **Biological Chemistry**, 2017. 398(5-6): p. 687-699.
187. Bantscheff, M., et al., *Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present.* **Analytical and Bioanalytical Chemistry**, 2012. 404(4): p. 939-965.
188. Zielinska, D.F., et al., *Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints.* **Cell**, 2010. 141(5): p. 897-907.
189. Nesvizhskii, A.I., *Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments.* **Proteomics**, 2012. 12(10): p. 1639-1655.
190. Choi, H., D. Fermin, and A.I. Nesvizhskii, *Significance analysis of spectral count data in label-free shotgun proteomics.* **Molecular & Cellular Proteomics**, 2008. 7(12): p. 2373-2385.
191. Karpievitch, Y.V., A.R. Dabney, and R.D. Smith, *Normalization and missing value imputation for label-free LC-MS analysis.* **BMC Bioinformatics**, 2012. 13(S16): p. S5.
192. Callister, S.J., et al., *Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics.* **Journal of Proteome Research**, 2006. 5(2): p. 277-286.
193. Zhang, B., et al., *Detecting differential and correlated protein expression in label-free shotgun proteomics.* **Journal of Proteome Research**, 2006. 5(11): p. 2909-2918.
194. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* **Nucleic Acids Research**, 1994. 22(22): p. 4673-4680.
195. Clamp, M., et al., *The Jalview Java alignment editor.* **Bioinformatics**, 2004. 20(3): p. 426-427.
196. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.* **Molecular Biology and Evolution**,

2013. 30(4): p. 772-780.
197. Camacho, C., et al., *BLAST+: architecture and applications.* **BMC Bioinformatics**, 2009. 10: p. 421.
198. UniProt, C., *The Universal Protein Resource (UniProt) in 2010.* **Nucleic Acids Research**, 2010. 38(Database issue): p. D142-D148.
199. Gotz, S., et al., *High-throughput functional annotation and data mining with the Blast2GO suite.* **Nucleic Acids Research**, 2008. 36(10): p. 3420-3435.
200. Jones, P., et al., *InterProScan 5: genome-scale protein function classification.* **Bioinformatics**, 2014. 30(9): p. 1236-1240.
201. Priebe, S., et al., *FungiFun2: a comprehensive online resource for systematic analysis of gene lists from fungal species.* **Bioinformatics**, 2015. 31(3): p. 445-446.
202. Lombard, V., et al., *The carbohydrate-active enzymes database (CAZy) in 2013.* **Nucleic Acids Research**, 2014. **42**(D1): p. D490-D495.
203. Horton, P., et al., *WoLF PSORT: protein localization predictor.* **Nucleic Acids Research**, 2007. 35(Web Server issue): p. W585-W587.
204. Käll, L., A. Krogh, and E.L. Sonnhammer, *Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server.* **Nucleic Acids Research**, 2007. 35(suppl_2): p. W429-W432.
205. Lum, G. and X.J. Min, *FunSecKB: the Fungal Secretome KnowledgeBase.* **Database** (Oxford), 2011. 2011: p. bar001.
206. Choi, J., et al., *Fungal secretome database: integrated platform for annotation of fungal secretomes.* **BMC Genomics**, 2010. 11: p. 105.
207. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* **Proceedings of the National Academy of Sciences**, 2005. 102(43): p. 15545-15550.
208. Chou, M.F. and D. Schwartz, *Biological sequence motif discovery using motif-x.* **Current Protocols in Bioinformatics**, 2011. 35(1): p. 13-15.
209. Amanchy, R., et al., *A curated compendium of phosphorylation motifs.* **Nature Biotechnology**, 2007. 25(3): p. 285-286.
210. Franceschini, A., et al., *STRING v9.1: protein-protein interaction networks, with increased coverage and integration.* **Nucleic Acids Research**, 2013. 41(D1): p. D808-D815.
211. Snel, B., et al., *STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene.* **Nucleic Acids Research**, 2000. 28(18): p. 3442-3444.
212. Shannon, P., et al., *Cytoscape: A software environment for integrated models of biomolecular interaction networks.* **Genome Research**, 2003. 13(11): p. 2498-2504.
213. Bader, G.D. and C.W. Hogue, *An automated method for finding molecular complexes*

*in large protein interaction networks.* **BMC Bioinformatics**, 2003. 4: p. 2.

214. Maere, S., K. Heymans, and M. Kuiper, *BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.* **Bioinformatics**, 2005. 21(16): p. 3448-3449.
215. Kolbusz, M.A., et al., *Transcriptome and exoproteome analysis of utilization of plant-derived biomass by Myceliophthora thermophila.* **Fungal Genetics and Biology**, 2014. 72: p. 10-20.
216. Gaskell, J., et al., *Transcriptome and secretome analyses of the wood decay fungus Wolfiporia cocos support alternative mechanisms of lignocellulose conversion.* **Applied and Environmental Microbiology**, 2016. 82(13): p. 3979-3987.
217. Vanden Wymelenberg, A., et al., *Comparative transcriptome and secretome analysis of wood decay fungi Postia placenta and Phanerochaete chrysosporium.* **Applied and Environmental Microbiology**, 2010. 76(11): p. 3599-3610.
218. Kuuskeri, J., et al., *Time-scale dynamics of proteome and transcriptome of the white-rot fungus Phlebia radiata: growth on spruce wood and decay effect on lignocellulose.* **Biotechnology for Biofuels and Bioproducts**, 2016. 9(1): p. 192.
219. Navarro, D., et al., *Fast solubilization of recalcitrant cellulosic biomass by the basidiomycete fungus Laetisaria arvalis involves successive secretion of oxidative and hydrolytic enzymes.* **Biotechnology for Biofuels and Bioproducts**, 2014. **7**(1): p. 143.
220. Schneider, W.D.H., et al., *Penicillium echinulatum secretome analysis reveals the fungi potential for degradation of lignocellulosic biomass.* **Biotechnology for Biofuels and Bioproducts**, 2016. 9: p. 66.
221. Benoit, I., et al., *Closely related fungi employ diverse enzymatic strategies to degrade plant biomass.* **Biotechnology for Biofuels and Bioproducts**, 2015. 8: p. 107.
222. Xiong, Y., et al., *The proteome and phosphoproteome of Neurospora crassa in response to cellulose, sucrose and carbon starvation.* **Fungal Genetics and Biology**, 2014. 72: p. 21-33.
223. Horta, M.A.C., et al., *Broad substrate-specific phosphorylation events are associated with the initial stage of plant cell wall recognition in Neurospora crassa.* **Frontiers in Microbiology**, 2019. 10: p. 2317.
224. Fiehn, O., *Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks.* **Comparative and Functional Genomics**, 2001. 2(3): p. 155-168.
225. Villas-Boas, S.G., et al., *Mass spectrometry in metabolome analysis.* **Mass Spectrometry Reviews**, 2005. **24**(5): p. 613-646.
226. Beale, D.J., et al., *Review of recent developments in GC-MS approaches to metabolomics-based research.* **Metabolomics**, 2018. 14(11): p. 152.

227. Ren, J.-L., et al., *Advances in mass spectrometry-based metabolomics for investigation of metabolites.* **RSC Advances**, 2018. 8(40): p. 22335-22350.
228. Chen, F., R. Ma, and X.-L. Chen, *Advances of metabolomics in fungal pathogen–plant interactions.* **Metabolites**, 2019. 9(8): p. 169.
229. Gummer, J.P., et al., *Metabolomics protocols for filamentous fungi*, in *In Plant Fungal Pathogens*. 2012, **Humana Press**. p. 237-254.
230. Markley, J.L., et al., *The future of NMR-based metabolomics.* **Current Opinion in Biotechnology**, 2017. 43: p. 34-40.
231. Wishart, D.S., *NMR metabolomics: A look ahead.* Journal of Magnetic **Resonance**, 2019. 306: p. 155-161.
232. Smedsgaard, J. and J. Nielsen, *Metabolite profiling of fungi and yeast: from phenotype to metabolome by MS and informatics.* **Journal of Experimental Botany**, 2005. 56(410): p. 273-286.
233. Roessner, U. and J. Bowne, *What is metabolomics all about?* **Biotechniques**, 2009. 46(5): p. 363-365.
234. Kluger, B., S. Lehner, and R. Schuhmacher, *Metabolomics and secondary metabolite profiling of filamentous fungi*, in *Biosynthesis and Molecular Genetics of Fungal Secondary Metabolites, Volume 2*. 2015, **Springer**, New York, NY. p. 81-101.
235. Ribbenstedt, A., H. Ziarrusta, and J.P. Benskin, *Development, characterization and comparisons of targeted and non-targeted metabolomics methods.* **PLoS One**, 2018. 13(11): p. e0207082.
236. Roberts, L.D., et al., *Targeted metabolomics.* **Current Protocols in Molecular Biology**, 2012. 98(1): p. 30-32.
237. Schuhmacher, R., et al., *Metabolomics and metabolite profiling.* **Analytical and Bioanalytical Chemistry**, 2013. 405(15): p. 5003-5004.
238. Pluskal, T., et al., *MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data.* **BMC Bioinformatics**, 2010. 11: p. 395.
239. Lommen, A. and H.J. Kools, *MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware.* **Metabolomics**, 2012. 8(4): p. 719-726.
240. Rost, H.L., et al., *OpenMS: a flexible open-source software platform for mass spectrometry data analysis.* **Nature Methods**, 2016. 13(9): p. 741-748.
241. Tautenhahn, R., et al., *XCMS Online: a web-based platform to process untargeted metabolomic data.* **Analytical Chemistry**, 2012. 84(11): p. 5035-5039.
242. Smith, C.A., et al., *METLIN: a metabolite mass spectral database.* **Therapeutic Drug Monitoring**, 2005. 27(6): p. 747-751.
243. Grossetete, S., B. Labedan, and O. Lespinet, *FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology.* **BMC Genomics**, 2010. 11: p. 81.
244. Kastenmuller, G., et al., *metaP-server: a web-based metabolomics data analysis*

*tool.* **Journal of Biomedicine and Biotechnology**, 2011. 2011.

245. Chavali, A.K. and S.Y. Rhee, *Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites.* **Briefings in Bioinformatics**, 2018. 19(5): p. 1022-1034.
246. Fedorova, N.D., V. Moktali, and M.H. Medema, *Bioinformatics approaches and software for detection of secondary metabolic gene clusters.* **Methods in Molecular Biology**, 2012. 944: p. 23-45.
247. Navarro-Munoz, J.C., et al., *A computational framework to explore large-scale biosynthetic diversity.* **Nature Chemical Biology**, 2020. 16(1): p. 60-68.
248. Khaldi, N., et al., *SMURF: Genomic mapping of fungal secondary metabolite clusters.* **Fungal Genetics and Biology**, 2010. 47(9): p. 736-741.
249. Umemura, M., et al., *MIDDAS-M: motif-independent de novo detection of secondary metabolite gene clusters through the integration of genome sequencing and transcriptome data.* **PloS One**, 2013. 8(12): p. e84028.
250. Wolf, T., et al., *CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes.* **Bioinformatics**, 2016. 32(8): p. 1138-1143.
251. Yi, G., S.H. Sze, and M.R. Thon, *Identifying clusters of functionally related genes in genomes.* **Bioinformatics**, 2007. 23(9): p. 1053-1060.
252. Röttig, M., et al., *NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity.* **Nucleic Acids Research**, 2011. 39(suppl_2): p. W362-W367.
253. Takeda, I., et al., *Motif-Independent Prediction of a Secondary Metabolism Gene Cluster Using Comparative Genomics: Application to Sequenced Genomes of Aspergillus and Ten Other Filamentous Fungal Species.* **DNA Research**, 2014. 21(4): p. 447-457.
254. Andersen, M.R., et al., *Accurate prediction of secondary metabolite gene clusters in filamentous fungi.* **Proceedings of the National Academy of Sciences**, 2013. 110(1): p. E99-E107.
255. Vesth, T.C., J. Brandl, and M.R. Andersen, *FunGeneClusterS: Predicting fungal gene clusters from genome and transcriptome data.* **Synthetic and Systems Biotechnology**, 2016. 1(2): p. 122-129.
256. Conway, K.R. and C.N. Boddy, *ClusterMine360: a database of microbial PKS/NRPS biosynthesis.* **Nucleic Acids Research**, 2012. 41(D1): p. D402-D407.
257. Hadjithomas, M., et al., *IMG-ABC: A knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites.* **mBio**, 2015. 6(4): p. e00932.
258. Medema, M.H., et al., *Minimum information about a biosynthetic gene cluster.* **Nature Chemical Biology**, 2015. 11(9): p. 625-631.
259. Peiris, D., et al., *Metabolite profiles of interacting mycelial fronts differ for pairings*

*of the wood decay basidiomycete fungus, Stereum hirsutum with its competitors Coprinus micaceus and Coprinus disseminatus.* **Metabolomics**, 2008. 4(1): p. 52.

260. Bertrand, S., et al., *Detection of metabolite induction in fungal co-cultures on solid media by high-throughput differential ultra-high pressure liquid chromatography-time-of-flight mass spectrometry fingerprinting.* **Journal of Chromatography A**, 2013. 1292: p. 219-228.
261. Chatterjee, S., et al., *Interactions among filamentous fungi Aspergillus niger, Fusarium verticillioides and Clonostachys rosea: fungal biomass, diversity of secreted metabolites and fumonisin production.* **BMC Microbiology**, 2016. 16(1): p. 83.
262. Luo, F., et al., *Metabolomic differential analysis of interspecific interactions among white rot fungi Trametes versicolor, Dichomitus squalens and Pleurotus ostreatus.* **Scientific Reports**, 2017. 7(1): p. 1-11.
263. Bhandari, D.R., et al., *Histology-guided high-resolution AP-SMALDI mass spectrometry imaging of wheat-Fusarium graminearum interaction at the root–shoot junction.* **Plant Methods**, 2018. 14(1): p. 103.
264. Chen, F., et al., *Combined metabonomic and quantitative RT-PCR analyses revealed metabolic reprogramming associated with Fusarium graminearum resistance in transgenic Arabidopsis thaliana.* **Frontiers in Plant Science**, 2018. 8: p. 2177.
265. Jacob, S., et al., *Unravelling the biosynthesis of pyriculol in the rice blast fungus Magnaporthe oryzae.* **Microbiology**, 2017. 163(4): p. 541-553.
266. Doehlemann, G., et al., *Reprogramming a maize plant: transcriptional and metabolic changes induced by the fungal biotroph Ustilago maydis.* **The Plant Journal**, 2008. 56(2): p. 181-195.
267. Hayden, H.L., et al., *Metabolomics approaches for the discrimination of disease suppressive soils for Rhizoctonia solani AG8 in cereal crops using 1H NMR and LC-MS.* **Science of The Total Environment**, 2019. 651: p. 1627-1638.
268. Verwaaijen, B., et al., *A comprehensive analysis of the Lactuca sativa, L. transcriptome during different stages of the compatible interaction with Rhizoctonia solani.* **Scientific Reports**, 2019. 9(1): p. 1-12.
269. Hu, Z., et al., *Metabolic profiling to identify the latent infection of strawberry by Botrytis cinerea.* **Evolutionary Bioinformatics**, 2019. 15: p. 1176934319838518.
270. Camañes, G., et al., *An untargeted global metabolomic analysis reveals the biochemical changes underlying basal resistance and priming in Solanum lycopersicum, and identifies 1-methyltryptophan as a metabolite involved in plant responses to Botrytis cinerea and Pseudomonas syringae.* **The Plant Journal**, 2015. 84(1): p. 125-139.
271. Negri, S., et al., *The induction of noble rot (Botrytis cinerea) infection during postharvest withering changes the metabolome of grapevine berries (Vitis vinifera*

*L., cv. Garganega).* **Frontiers in Plant Science**, 2017. 8: p. 1002.

272. Ghosh, S., et al., *Proteometabolomic analysis of transgenic tomato overexpressing oxalate decarboxylase uncovers novel proteins potentially involved in defense mechanism against Sclerotinia.* **Journal of Proteomics**, 2016. 143: p. 242-253.
273. Lee, S., et al., *Targeted metabolomics for Aspergillus oryzae-mediated biotransformation of soybean isoflavones, showing variations in primary metabolites.* **Bioscience, Biotechnology, and Biochemistry**, 2014. 78(1): p. 167-174.
274. Tao, N., et al., *The terpene limonene induced the green mold of citrus fruit through regulation of reactive oxygen species (ROS) homeostasis in Penicillium digitatum spores.* **Food Chemistry**, 2019. 277: p. 414-422.
275. Xiong, Y., et al., *A fungal transcription factor essential for starch degradation affects integration of carbon and nitrogen metabolism.* **PLOS Genetics**, 2017. 13(5): p. e1006737.
276. Daly, P., et al., *Colonies of the fungus Aspergillus niger are highly differentiated to adapt to local carbon source variation.* **Environmental Microbiology**, 2020. 22(3): p. 1154-1166.
277. Huang, S., K. Chaudhary, and L.X. Garmire, *More is better: recent progress in multi-omics data integration methods.* **Frontiers in Genetics**, 2017. 8: p. 84.
278. Schaepe, P., et al., *Updating genome annotation for the microbial cell factory Aspergillus niger using gene co-expression networks.* **Nucleic Acids Research**, 2019. 47(2): p. 559-569.
279. Martinez, D., et al., *Genome, transcriptome, and secretome analysis of wood decay fungus Postia placenta supports unique mechanisms of lignocellulose conversion.* **Proceedings of the National Academy of Sciences**, 2009. 106(6): p. 1954-1959.
280. Daly, P., et al., *Dichomitus squalens partially tailors its molecular responses to the composition of solid wood.* **Environmental Microbiology**, 2018. 20(11): p. 4141-4156.
281. Samal, A., et al., *Network reconstruction and systems analysis of plant cell wall deconstruction by Neurospora crassa.* **Biotechnology for Biofuels and Bioproducts**, 2017. 10: p. 225.
282. Brandl, J., et al., *A community-driven reconstruction of the Aspergillus niger metabolic network.* **Fungal Biology and Biotechnology**, 2018. 5: p. 16.
283. Wanichthanarak, K., J.F. Fahrmann, and D. Grapov, *Genomic, proteomic, and metabolomic data integration strategies.* **Biomarker Insights**, 2015. 10: p. BMI-S29511.
284. Lê Cao, K.-A., I. González, and S. Déjean, *integrOmics: an R package to unravel relationships between two omics datasets.* **Bioinformatics**, 2009. 25(21): p. 2855-2856.

285. Tuncbag, N., et al., *SteinerNet: a web server for integrating 'omic'data to discover hidden components of response pathways.* **Nucleic Acids Research**, 2012. **40**(W1): p. W505-W509.
286. Tuncbag, N., et al., *Network-based interpretation of diverse high-throughput datasets through the omics integrator software package.* **PLoS Computational Biology**, 2016. 12(4).
287. Rohart, F., et al., *mixOmics: An R package for 'omics feature selection and multiple data integration.* **PLoS Computational Biology**, 2017. 13(11): p. e1005752.
288. Wang, B., et al., *Similarity network fusion for aggregating data types on a genomic scale.* **Nature Methods**, 2014. 11(3): p. 333-337.
289. Shen, R., A.B. Olshen, and M. Ladanyi, *Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis.* **Bioinformatics**, 2009. 25(22): p. 2906-2912.
290. Meng, C., et al., *A multivariate approach to the integration of multi-omics datasets.* **BMC Bioinformatics**, 2014. 15: p. 162.
291. Ramazzotti, D., et al., *Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival.* **Nature Communications**, 2018. 9(1): p. 1-14.
292. Surujon, D. and T. van Opijnen, *ShinyOmics: collaborative exploration of omics-data.* **BMC Bioinformatics**, 2020. 21(1): p. 22.
293. Miyauchi, S., et al., *The integrative omics of white-rot fungus Pycnoporus coccineus reveals co-regulated CAZymes for orchestrated lignocellulose breakdown.* **PLoS One**, 2017. 12(4): p. e0175528.
294. Swift, C.L., et al., *Linking "omics" to function unlocks the biotech potential of non-model fungi.* **Current Opinion in Systems Biology**, 2019. 14: p. 9-17.
295. Aguilar-Pontes, M.V., et al., *The gold-standard genome of Aspergillus niger NRRL 3 enables a detailed view of the diversity of sugar catabolism in fungi.* **Studies in Mycology**, 2018. 91: p. 61-78.
296. Otto, T.D., et al., *RATT: rapid annotation transfer tool.* **Nucleic Acids Research**, 2011. 39(9): p. e57.

# Chapter 3

# The sugar metabolic model of *Aspergillus niger* can only be reliable transferred to fungi of its phylum

This chapter was published in *Journal of Fungi*

Jiajia Li, Tania Chroumpi, Sandra Garrigues, Roland S. Kun, Jiali Meng, Sonia Salazar-Cerezo, Maria Victoria Aguilar-Pontes, Yu Zhang, Sravanthi Tejomurthula, Anna Lipzen, Vivian Ng, Chaevien S. Clendinen, Nikola Tolić, Igor Grigoriev, Adrian Tsang, Miia R. Mäkelä, Berend Snel, Mao Peng, Ronald P. de Vries

Volume 8, Pages 1315, December 2022

DOI: https://doi.org/10.3390/jof8121315

## Abstract

Fungi play a critical role in the global carbon cycle by degrading plant polysaccharides to small sugars and metabolizing them as carbon and energy sources. We mapped the well-established sugar metabolic network of *Aspergillus niger* to five taxonomically distant species (*Aspergillus nidulans*, *Penicillium subrubescens*, *Trichoderma reesei*, *Phanerochaete chrysosporium*, and *Dichomitus squalens*) using an orthology-based approach.

The diversity of sugar metabolism correlates well with the taxonomic distance of the fungi. The pathways are highly conserved between the three studied Eurotiomycetes (*A. niger, A. nidulans*, *P. subrubescens*). A higher level of diversity was observed between the *T. reesei* and *A. niger*, and even more so for the two Basidiomycetes. These results were confirmed by integrative analysis of transcriptome, proteome and metabolome, as well as growth profiles of the fungi growing on the corresponding sugars.

In conclusion, the establishment of sugar pathway models in different fungi revealed the diversity of fungal sugar conversion and provided a valuable resource for the community, which would facilitate rational metabolic engineering of these fungi as microbial cell factories.

## 1. Introduction

Plant biomass is one of the most abundant renewable resources on earth, and it serves as a primary raw material in a variety of industries [1]. In nature, fungi can effectively release monosaccharides from plant biomass polysaccharides [2-4], and further metabolize them for energy and growth through a variety of sugar catabolic pathways [5]. D-glucose is by far the most abundant monosaccharide in plant biomass, and it is the major component of cellulose, starch and some hemicelluloses. D-fructose is found as a free monosaccharide in many fruits and plants, as well as being a major component of the disaccharide sucrose and polymer inulin [6]. Typically, D-glucose and D-fructose are easily taken up from the environment and catabolized by fungi, making them preferential monomeric carbon sources. D-fructose enters the cell and is phosphorylated to fructose 6-phosphate before entering glycolysis. D-glucose is phosphorylated to glucose 6-phosphate by glucokinase [7] and hexokinase [8], which can then be converted into fructose 6-phosphate. The pentoses D-xylose and L-arabinose are converted through the pentose catabolic pathway (PCP) to D-xylulose-5-phosphate, which enters the pentose phosphate pathway (PPP) [9]. Other sugars, like D-galactose, D-mannose, D-galacturonic acid, L-rhamnose and D-gluconic acid are converted through sugar-specific metabolic pathways [5, 10] and eventually all entering glycolysis. The final product of glycolysis is further metabolized through the tricarboxylic acid (TCA) and the glyoxylate cycles. The TCA cycle is present in almost all aerobic living organisms, including most fungi [11, 12]. The glyoxylate cycle is essentially a truncated version of the TCA, which metabolizes acetyl-CoA without the loss of $CO_2$ [13]. The inputs to both cycles are acetyl-CoA and oxaloacetate, which originate from pyruvate of glycolysis. In short, fungal central carbon catabolism is a complex network that involves many pathways.

Sugar catabolism by fungi is directly linked to their ability to utilize plant biomass as sugars are the main components of these substrates. A systematical

investigation of the components and diversity of sugar metabolic networks of different species is crucial for our understanding of the role of fungi in their natural environment as well as many fungal industrial applications. In a previous study of *A. niger* NRRL 3 [10], a model for the sugar metabolic pathways of this species has been established, which consists of 155 genes involved in 11 sugar metabolic pathways (glycolysis, PCP, PPP, D-galactose metabolism, D-mannose metabolism, D-gluconate metabolism, D-galacturonic acid catabolism, L-rhamnose metabolism, glycerol metabolism, TCA and glyoxylate cycles). Meanwhile, recent experimental studies have refined this predicted sugar metabolic network and identified new enzymes involved in L-rhamnose catabolism [14] and PCP [9]. In this study, we aimed to transfer the sugar metabolic model of *A. niger* to five taxonomically diverse fungi (*Aspergillus nidulans*, *Penicillium subrubescens*, *Trichoderma reesei*, *Phanerochaete chrysosporium*, and *Dichomitus squalens*) using an orthology-based approach. In addition, we filtered out the potential pseudogenes and compared the expression profiles of the predicted enzyme-encoding genes based on transcriptomic, proteomic and metabolomic data of fungi grown on diverse sugars, and partially validated the predicted sugar catabolic pathways with fungal growth profiles on the corresponding sugars.

## 2. Materials and Methods

### *2.1 Fungal strains*

The five fungi chosen for this study differ in their taxonomic distance to *A. niger*, ranging from very close (*A. nidulans* - a member of a different clade of the same genus) to close (*P. subrubescens* - sister genus in the order of Eurotiales) to more distant (*T. reesei* - belonging to the Sordariomycetes, a different ascomycete class), to very distant (*P. chrysosporium* and *D. squalens* - members of a different phylum, Basidiomycota). The details of the fungal species of this study are denoted in Table S1.

### *2.2 Identification of orthologs of sugar metabolism genes across six fungal species*

Orthologous SMGs were identified using OrthoMCL (https://orthomcl.org/orthomcl/app) [15] and OrthoFinder (https://github.com/davidemms/OrthoFinder) [16]. The protein sequences of the fungal species were downloaded from the JGI MycoCosm Portal (https://mycocosm.jgi.doe.gov) [17] as input for these two tools. For OrthoMCL, the protein sequences of the selected fungal genomes were compared using all-against-all BLASTP with a P-value cut-off of 1e-5. Many-to-many ortholog groups including recent paralogs were then detected based on all-against-all BLAST searches of complete proteomes as described previously [15]. The OrthoFinder method provides a fast, accurate and comprehensive platform to infer the complete set of orthologs between selected species based on the phylogenetic information from the ortho-group tree. OrthoFinder was performed using default parameters with DIAMOND [18] for sequence similarity searches and DendroBLAST [19] for the tree inference of orthogroups. Since the OrthoFinder method is an alternative and auxiliary method to OrthoMCL, it can predict orthologs that are misidentified by OrthoMCL. We merged ortholog mapping results of these two tools as our final ortholog set. For any two ortholog groups identified by OrthoMCL and OrthoFinder that share one or more common member genes, we merged the two ortholog groups into a new group by taking the combined set of all their member genes.



### *2.3 Data gathering and preparation for Pathway Tools*

The sugar metabolic pathway construction was carried out by the PathoLogic component of Pathway Tools v24.0 software (http://pathwaytools.com/) [20, 21]. This software enables users to generate and manage an interactively edited organism-specific database (DB) named Pathway Genome Database (PGDB). The PGDB is built from an annotated genome of an organism, which integrates information of the genome, genes, proteins, biochemical reactions, and predicted metabolic pathways and operons. The annotated genome of

an organism was imported into PathoLogic to predict metabolic pathways. We collected annotation files derived from a wide variety of sources. The Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.kegg.jp/) annotation [22], Gene Ontology (GO) annotation [23], InterPro database annotation [24], SignalP annotation [25], Eukaryotic orthologous groups (KOG) annotation [26], Pfam [27, 28] and genome functional information (.gff file) were all derived from the JGI MycoCosm Portal (https://mycocosm.jgi.doe.gov) [17]. In addition, we included the annotation of best BLAST hit in the Swiss-Prot [29] database with an E-value <1e−5, as well as functional annotation obtained from our ortholog mapping analysis (Table S2). To combine all these annotations, we created a home-made Perl pipeline to generate a set of PathoLogic format files that can be recognized by the PathoLogic component of Pathway Tools software to automatically build databases. After uploading the annotation files for PathoLogic, the program can infer metabolic pathways and enzymes by assessing the genome annotation with respect to a series of reference databases of metabolic pathways, such as EcoCyc [30] and MetaCyc databases [31]. Finally, the metabolic pathways were visualized in the corresponding organism-specific DBs, and we manually refined them by filling gaps according to ortholog mapping results and removing the predicted genes sharing low sequence homology to the reference gene of *A. niger* and excluding potential pseudogenes with extremely low expression levels (maximum FPKM < 10 in the transcriptomes of all tested conditions).

### *2.4 Transcriptome sequencing and analysis*

The transcriptome data of *A. niger* [32] and *D. squalens* [33] were obtained from our previous studies (Gene Expression Omnibus (GEO) database accessions: GSE98572 and GSE105076). Transcriptome data of the other four species grown on nine monosaccharides were newly generated in this study. In detail, the *A. nidulans* FGSC A4 and *P. subrubescens* FBCC1632/CBS132785 were pre-cultured in complete medium [34] with 2% D-fructose and mycelial aliquots were then transferred to minimal medium [34] with 25

mM D-glucose, D-fructose, D-galactose, D-mannose, L-rhamnose, D-xylose, L-arabinose, D-galacturonic acid or D-glucuronic acid, respectively, and cultivated for 2 h. The same cultivation approach was used for *T. reesei* QM6a and *P. chrysosporium* PR-78, but with media optimized for these species [35, 36]. Mycelial samples were harvested after 4 h and immediately frozen in liquid nitrogen.

Total RNA was extracted from ground mycelial samples using TRIzol reagent (Invitrogen) according to the instructions of the manufacturer. Purification of mRNA, synthesis of cDNA library and sequencing were conducted at DOE Joint Genome Institute (JGI) as previously described [14]. The reads from each of the transcriptome sequencing (RNA-seq) samples were deposited in the Sequence Read Archive at NCBI under the accession numbers: *A. nidulans* SRP262827-SRP262853, *P. subrubescens* SRP246823-SRP246849, *T. reesei* SRP378720-SRP378745, and *P. chrysosporium* SRP249214-SRP249240.

In this study, we filtered out the genes whose maximum expression levels (FPKM) were less than 10 in all tested conditions.

### *2.5 Proteome quantitation and statistical analysis*

The sample preparation and proteomics analysis of intracellular proteins was performed similarly as previously described [37]. Briefly, the intracellular proteome was analyzed using equal amounts of proteins from each sample. MS analysis was performed using a Q-Exactive Plus mass spectrometer (Thermo Scientific) outfitted with a homemade nano-electrospray ionization interface. The ion transfer tube temperature and spray voltage were 300°C and 2.2 kV, respectively. Data were collected for 120 min following a 10 min delay after completion of sample trapping and start of gradient. FT-MS spectra were acquired from 300 to 1800 m/z at a resolution of 70 k (AGC target 3e6) and the top 12 FT-HCD-MS/MS spectra were acquired in data-dependent mode with an isolation window of 1.5 m/z at a resolution of 17.5 k (AGC target 1e5) using a normalized collision energy of 30, dynamic exclusion

time of 30 s , and detected charge state of an ion 2 to 8. Generated MS/MS spectra were searched against protein sequences of each fungus obtained from MycoCosm (Table S1) using (MSGF+) [38, 39]. Best matches from the MSGF+ searches were filtered at 1% FDR and MASIC software [40] was used to pull abundances for identified peptides. Only protein specific peptides (peptides unique to protein in the whole protein collection) were used in consequent analysis and aggregation. InfernoR software [41] was used to transform peptides abundances (log2) and perform mean central tendency normalization. Protein grouped normalized peptide abundances were de-logged, summed, transformed (log2) and normalized again in InfernoR to produce normalized abundances for the protein level roll-up (Table S5). Proteins abundances were then filled with zeros for missing values. The average values of nonzero abundances from the three replicates of each carbon source were used to represent the corresponding protein abundance of each growth condition. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the MassIVE partner repository with the data set identifier (MSV000090477).

### *2.6 Metabolomics data preprocessing and analysis*

Metabolomics analysis of intracellular metabolites was performed as previously described [42]. Briefly, metabolites were extracted from the ground mycelia underwent methoximation and silylation with N-methyl-N-trimethylsilyltrifluoroacetamide and 1% trimethylchlorosilane (MSTFA). Derivatized samples were analyzed using an Agilent GC 7890A coupled with a single quadrupole MSD 5975C (Agilent Technologies) with a standard mixture of fatty acid methyl esters (FAMEs) for RI alignment. GCMS raw files were processed and analyzed using Metabolite Detector software, version 2.5 beta [43]. The GCMS data have been deposited to the MassIVE database with accession number MSV000090441. A correlation analysis between the expression values of SMGs and the abundances of related metabolites was performed by using WGCNA R package [44]. Only significant correlation

between SMGs and metabolites (with the absolute Pearson correlation coefficient (PCC) ≥ 0.5 and *p*-value < 0.05) were selected for further analysis.

### *2.7 Growth profiling on nine monosaccharides*

For growth phenotype analyses, strains were grown on minimal medium (MM) [34], with the exception of *T. reesei* and *P. chrysosporium* that were grown on optimized medium for these species [35, 36], on 1.5% (w/v) agar plates with one of nine monosaccharides (http:/www.fung-growth.org/), including 25 mM D-glucose, D-fructose, D-galactose, D-mannose, L-rhamnose, D-xylose, L-arabinose, D-galacturonic acid and D-glucuronic acid. Growth was performed at 30°C for both Aspergilli and 25°C for the other species. Media with no carbon source was used as a control. If growth on a specific carbon source is the same as with no carbon source, it is considered as no growth.



## 3. Results

### *3.1 Identification of orthologs between A. niger and five other fungi*

Orthologs are genes in different species that originated from a common ancestor after a speciation event, resulting in the retention of similar functions during evolution. An orthology-based approach can assist us in inferring functions among different organisms. We identified orthologs among three Eurotiomycota species (*A. niger, A. nidulans* and *P. subrubescens*), one Sordariomycota species (*T. reesei*) and two Basidiomycota species (*P. chrysosporium* and *D. squalens*) by using OrthoMCL and OrthoFinder (see Materials and Methods). Except for one gene involved in glyoxylate cycle and one gene involved in D-gluconate metabolism, 153 sugar metabolism-related genes (SMGs) of *A. niger* had one or more copies in at least one of five tested fungi (Table S2). Subsequently, the ortholog mapping results were incorporated into PathoLogic files to build the sugar metabolic networks.

### *3.2 Generation of sugar metabolic models for the selected fungal species*

We established the sugar metabolic pathways for each of the selected fungi by integrating diverse annotation sources using the Pathway Tools software [20, 21]. Furthermore, we excluded candidate genes with extremely low expression levels (see Materials and Methods, Table S3 and Table S4). As shown in Figure 1, we summarized the total number of genes predicted to be involved in each specific sugar metabolism pathway of each studied fungus, as well as the overall completeness of the pathways based on the percentage of total reactions predicted for each pathway of each studied fungus compared to the reactions reported for *A. niger*. Overall, most of sugar metabolic pathways and SMGs are highly conserved between *A. niger* and the studied fungi, while the galactose oxidoreductive pathway and D-gluconate metabolic pathways showed clear variations among the studied species, and the glycerol, L-rhamnose and PCP pathways particularly differed between the Ascomycota and Basidiomycota species (Figure 1). In the following sections we will detail the evolutionary conservation and multi-omics profiles of the predicted sugar metabolism across different species, as well as the link between the predicted sugar metabolic network and growth profile.

### *3.3 Conservation of sugar metabolism among different fungal species*

#### *3.3.1 Sugar catabolic pathways with strong conservation*

The sugar metabolic models revealed that glycolysis, TCA and glyoxylate cycles, and D-galacturonic acid catabolic pathway are highly conserved among the tested fungi (Figure 1).

Glycolysis

Glycolysis is one of the major pathways of central metabolism and plays a key role in the growth of almost all organisms [45]. It converts glucose into pyruvate along with the formation of adenosine triphosphate (ATP) and nicotinamide adenine dinucleotide (NADH).

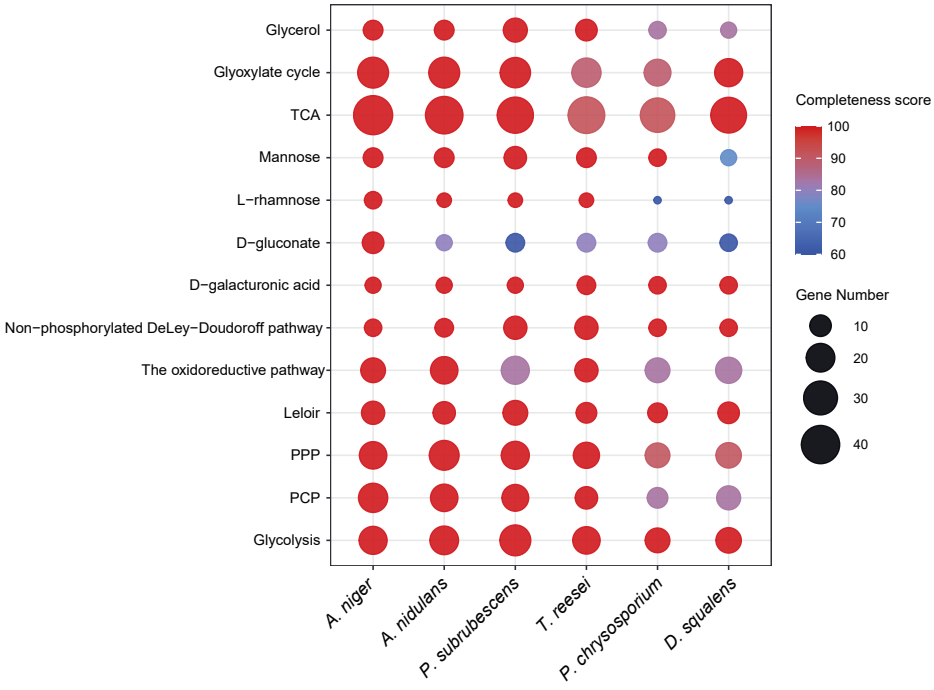


**Figure 1. Conservation of sugar metabolic pathways among *A. niger*, *A. nidulans*, *T. reesei*, *P. subrubescens*, *P. chrysosporium* and *D. squalens*.** The size of the dots indicates the number of genes involved in each pathway, and the color indicates the completeness of the pathway. The completeness score of a pathway is defined as the percentage of predicted reactions in a studied fungus compared to the total reactions reported in *A. niger* for each specific sugar metabolic pathway.

Typically, fungal glycolysis is a sequence of ten enzymatic reactions catalyzed by the following enzymes, namely, glucokinase (GLK, EC 2.7.1.2) [7] and hexokinase (HXK, EC 2.7.1.1) [8], glucose 6-phosphate isomerase (PGI, EC 5.3.1.9) [46], phosphofructokinase (PFK, EC 2.7.1.11) [47], D-fructose 1,6-phosphatase (FBP, EC 3.1.3.11) [48], fructose-bisphosphate aldolase (FBA, EC 4.1.2.13) [49], triphosphate isomerase (TPI, EC 5.3.1.1), glyceraldehyde 3-phosphate dehydrogenase (GPD, EC 1.2.1.12) [50], phosphoglycerate kinase (PGK, EC 2.7.2.3) [51], phosphoglycerate mutase (PGM, EC 5.4.2.11/5.4.2.12) [31], enolase (ENO, EC 4.2.1.11) [52], as well as pyruvate kinase (PKI, EC 2.7.1.40) [53]. Additionally, the metabolism of

D-fructose is linked to glycolysis through hexokinase (HXK, EC 2.7.1.4) that converts D-fructose into D-fructose 6-phosphate [8, 54].

Given the essential role of the glycolysis pathway, it is not surprising that all the glycolytic enzyme-encoding genes were detected in all the six studied fungi (Figure 2, Table S3). In contrast to the completeness of glycolysis in all studied fungi, the gene copy numbers for different enzymes showed variations across different species. The PGI, PFK and PGK encoding genes are extremely conserved, and present as a single copy in each fungus, while the copy numbers of other genes show a slight difference between *A. niger* and the other species. Overall, the studied Ascomycetes seem to have more similar numbers of glycolysis-related genes than the two Basidiomycetes (Table S3). For instance, the glycolytic genes between *A. nidulans* and *A. niger* are very conserved, with almost the same numbers of orthologs identified for each enzymatic reaction step, except two more predicted genes encoding HXK/GLK and one fewer gene encoding PGM. *P. subrubescens* has more copies for both HXK/GLK and FBA. *T. reesei* has one extra copy for both HXK/GLK and PKI, but one copy less for FBA, GPD and PGM. In contrast, the two Basidiomycete species only have a single copy for HXK, GLK, PGI, PFK, FBP, TPI, GPD, PGK and PKI encoding genes. However, we identified one extra copy of ENO and PGM for *P. chrysosporium*, whereas in *D. squalens*, we found extra copies of FBA and PGM. In addition, many of our predicted genes related to glycolysis in *A. nidulans* and *T. reesei* have been experimentally characterized in previous studies [55, 56].

The tricarboxylic acid (TCA) and glyoxylate cycles

Similar to glycolysis, the TCA cycle is another major pathway of central metabolism, which is essential for energy metabolism in fungi [11, 12].

The input to the TCA cycle is acetyl-CoA, derived from metabolism of carbohydrates, fatty acids and amino acids. The TCA cycle is a closed loop consisting of eight enzymatic reactions, where the oxaloacetate reacting with acetyl-CoA in the first step is reformed in the last step (Figure 2). A common

source of acetyl-CoA is from the oxidation of pyruvate, the end product of glycolysis. In addition, acetyl-CoA is also provided from coenzyme A and acetate through the reaction catalyzed by acetyl-CoA synthetase (AcuA, EC 6.2.1.1) [57]. Acetate used in the above reaction can be obtained from the conversion of oxaloacetate into oxalate by oxaloacetate acetylhydrolase (OAH, EC 3.7.1.1) [58]. As a modified TCA cycle, the glyoxylate cycle shares three enzymatic reactions with the TCA cycle: malate dehydrogenase (MDH, EC 1.1.1.37), aconitate hydratase (ACO, EC 4.2.1.3), and citrate synthase (CIT, EC 2.3.3.1/2.3.3.16). Therefore, the glyoxylate cycle can access the reactions of the conversion of pyruvate to oxaloacetate and acetyl-CoA, as well as the other reactions of the conversion between oxaloacetate and acetyl-CoA through the intermediate compound acetate, as described in Figure 2.



We identified almost all the necessary genes for TCA and glyoxylate cycles in the genomes of all our tested fungi, except for the OAH encoding genes that are not present in *T. reesei* and *P. chrysosporium*. Similar to glycolysis, some of the genes are present in more than one copy (Table S3). For example, the *acuA* gene encoding acetyl-CoA synthetase has a single copy in Basidiomycetes, but has additional in-paralogs in Ascomycetes. The *acuD* gene encoding isocitrate lyase, which breaks down isocitrate into glyoxylate and succinate, has additional in-paralogs in all tested fungi except for *A. niger*. Two copies of pyruvate carboxylase (PYC, EC 6.4.1.1) were identified in the other two Eurotiomycetes, while one copy was found in *T. reesei* and the two Basidiomycetes.

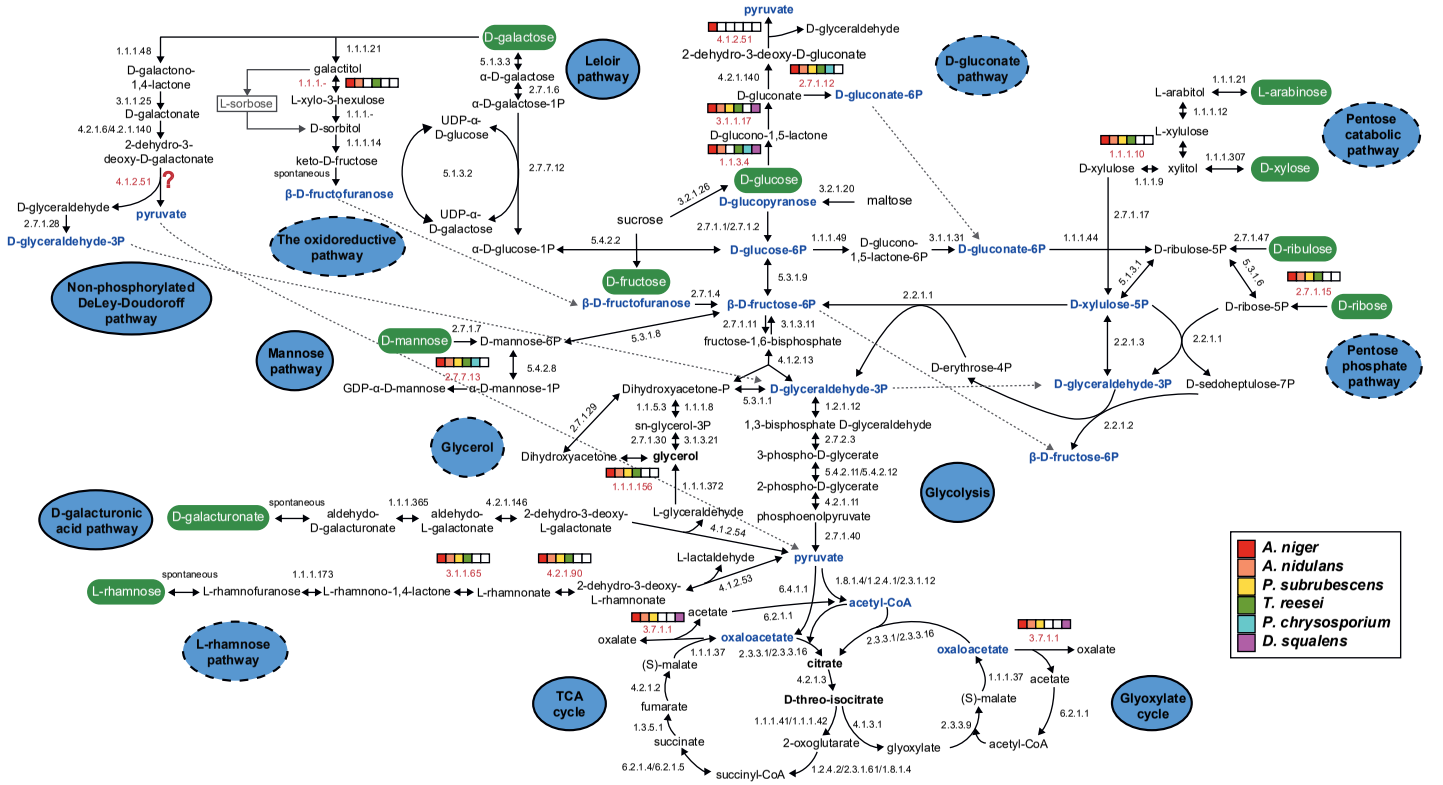


**Figure 2. Sugar metabolic networks of *A. niger*, *A. nidulans*, *T. reesei*, *P. subrubescens*, *P. chrysosporium* and *D. squalens*.** The names of the sugar metabolic pathways are shown in the blue circles. The solid or dashed lines of the circles indicate whether the corresponding pathway is complete or not, respectively. For enzymes involved in each reaction of each pathway, the corresponding enzyme EC number is shown, and for the reaction with the absence of genes in one or more of our studied species, the EC number is highlighted in red.

**(Figure 2.** Continued**)** A color bar was also used to highlight the presence and absence of genes in specific reactions in the studied species. The question marks indicate that the corresponding reaction remains to be discovered.

D-Galacturonic acid metabolism

D-Galacturonic acid is the main constituent of pectin, and it is an important carbon source for microorganisms living on plant material [59]. In fungi, D-galacturonic acid is catabolized through a non-phosphorylating pathway involved in five steps (Figure 2). After the spontaneous conversion of D-galacturonic acid to aldehyde-D-galacturonate, the latter is eventually converted into pyruvate and L-glyceraldehyde through three enzymes: D-galacturonic acid reductase (GaaA, EC 1.1.1.365), L-galactonate dehydratase (GaaB, EC 4.2.1.146), and 2-keto-3-deoxy-L-galactonate aldolase (GaaC, EC 4.1.2.54). The further conversion of L-glyceraldehyde to glycerol is catalyzed by L-glyceraldehyde reductase (GaaD/LarA, 1.1.1.372), which enters the glycerol catabolism [59-65].

Based on our prediction, the orthologous genes of all characterized enzymes in *A. niger* were identified in all five studied species (Table S3). Only small variations in the number of gene copy numbers across different species were found for the D-galacturonic acid pathway. The GaaA and GaaD/LarA encoding genes are extremely conserved, and are present as a single copy in each fungus. In addition, an exclusive orthologous gene encoding GaaB can be identified in the other four studied species, except for two copies in *T. reesei.* In addition, gene encoding GaaC only has a single copy in the two Eurotiomycetes but has two copies in *T. reesei* and in the two Basidiomycota species. Overall, this five-step D-galacturonic acid catabolism pathway is evolutionarily conserved in all the studied fungi (Figure 1).

*3.3.2 Sugar catabolism with moderate conservation*

Besides the highly conserved sugar catabolic pathways described above, several sugar catabolic pathways showed moderate conservation, which include glycerol metabolism, D-mannose metabolism, pentose phosphate

pathway and D-galactose catabolism. For each of these pathways, we failed to identify corresponding enzyme encoding genes for one or more essential reactions in at least one studied species.

Glycerol metabolism

Glycerol is an ubiquitous organic compound in nature that can be metabolized by many fungi as a source of carbon and energy [66, 67]. In addition, glycerol can be derived from the D-galacturonic acid catabolism pathway as described in the section above. Before glycerol enters glycolysis, it needs to be converted into dihydroxyacetone phosphate as shown in Figure 2.

The reversible conversion from glycerol to *sn*-glycerol 3-phosphate is mediated by glycerol kinase (GlcA, EC 2.7.1.30) and glycerol 1-phosphatase (GPP, EC 3.1.3.21). The further conversion of *sn*-glycerol 3-phosphate to dihydroxyacetone phosphate, which is also produced in glycolysis, is catalyzed by two different glycerol 3-phosphate dehydrogenases (GFD, EC 1.1.1.8; EC 1.1.5.3). In addition, glycerol can be converted to dihydroxyacetone and then to dihydroxyacetone phosphate by glycerol dehydrogenase (GldB, EC 1.1.1.156) and dihydroxyacetone kinase (DakA, EC 2.7.1.29), respectively [68-70].

Genomes of all the tested fungi encode at least one copy of each enzyme involved in this pathway, except for the two Basidiomycetes that lack *gldB*, which is required for reduction of glycerol to dihydroxyacetone (Figure 2). The *glcA* gene had additional in-paralogs (paralogs that arose after the species split) in *A. nidulans* and *P. subrubescens*. In addition, we found multiple copies of *gldB* in *P. subrubescens*, *T. reesei* and *P. chrysosporium,* which was present as a single copy in *A. niger* and *A. nidulans*.

D-Mannose metabolism

D-Mannose forms the backbone of polysaccharide mannan or galactomannan, which is one of the major constituents of hemicellulose in the plant cell wall [3]. Within the cell, D-mannose is phosphorylated by hexokinase (HXK,

EC 2.7.1.7) to produce D-mannose 6-phosphate. The latter can either be catabolized by mannose 6-phosphate isomerase (PMI, EC 5.3.1.8) into D-fructose 6-phosphate entering into glycolysis, or be converted to GDP-α-D-mannose through sequential actions of phosphomannomutase (PMM, EC 5.4.2.8) and mannose-1-phosphate guanylyltransferase (MGT, EC 2.7.7.13).

In *A. niger*, the formation of D-mannose 6-phosphate from D-mannose recruits the same *hxk* gene set involved in the conversion of D-glucose to D-glucose 6-phosphate [5, 10]. However, in *T. reesei* and the two Basidiomycetes, only parts of *hxk* genes were predicted to be involved in D-mannose catabolism (Table S3). In general, all reactions of mannose metabolism have at least one gene identified in the studied fungi, with the exception of genes involved in the conversion of α-D-mannose 1-phosphate to GDP-α-D-mannose missing in *D. squalens*.



Pentose phosphate pathway (PPP)

The pentose phosphate pathway (PPP) is an important part of central metabolism as well. It is the major source of NADPH, and also provides the intermediate ribose 5-phosphate that is essential for the synthesis of nucleotides and nucleic acids [71].

The PPP is connected with D-ribulose, D-ribose and D-xylulose catabolism and produces several glycolytic intermediates, including glucose 6-phosphate, D-fructose 6-phosphate, D-glyceraldehyde-3-phosphate, and NADPH that are essential components during glycolysis. The PPP can be divided into two phases, the oxidative phase and the non-oxidative phase. In the oxidative phase, glucose 6-phosphate is oxidized via three enzymatic steps to ribulose 5-phosphate with the generation of NADPH [72]. The first three reactions are catalyzed by glucose 6-phosphate dehydrogenase (GSD, EC 1.1.1.49), 6-phosphogluconolactonase (PGL, EC 3.1.1.31), and 6-phosphogluconate dehydrogenase (GND, EC 1.1.1.44). The non-oxidative branch is composed of a series of rearrangements catalyzed by an isomerase (ribose 5-phosphate isomerase, RPI, EC 5.3.1.6), an epimerase (ribulose-phosphate 3-epimerase,

RPE, EC 5.1.3.1), two different transketolases (TktA, EC 2.2.1.1/TktC, EC 2.2.1.3), and a transaldolase (TAL, EC 2.2.1.2), successively involving C5 compounds (D-xylulose 5-phosphate and D-ribose 5-phosphate), C3 and C7 compounds (D-glyceraldehyde 3-phosphate and D-sedoheptulose 7-phosphate), and C4 and C6 compounds ( D-erythrose 4-phosphate and D-fructofuranose 6-phosphate). In addition, the conversion between D-erythrose 4-phosphate and D-glyceraldehyde 3-phosphate can also be catalyzed by TktA. Unlike the oxidative branch, the non-oxidative branch is reversible. Another source of D-ribulose 5-phosphate and D-ribose 5-phosphate is derived from the phosphorylation of D-ribulose and D-ribose by ribulokinase (RBT, EC 2.7.1.47) and ribokinase (RBK, EC 2.7.1.15), respectively.

Except for the absence of RBK in the two studied Basidiomycota species, all reactions in the PPP have their corresponding enzymes assigned (Figure 2 and Table S3). However, the number of genes involved in specific reactions of different fungi shows large variability (Table S3). For example, the genes encoding GsdA and RpeA are extremely conserved. The former is present as a single copy in each fungus, and the latter has a single copy in five fungi apart from *P. subrubescens* possessing two copies. The PglA encoding gene has a single copy in the three studied Eurotiomycetes, while additional copies were observed in other fungi. The other genes showed different patterns across different fungi. For instance, RPI encoding gene has additional copies in the three Eurotiomycetes, but a single copy in the one Sordariomycete and the two Basidiomycetes. The two TKT enzymes showed different conservation among different species. The TktC encoding gene has a single copy in five studied fungi except for *A. niger* that has one more copy. The TktA-encoding gene has a single copy in *A. niger* and *T. reesei*, but two copies in the other fungi. RbtA-encoding gene is present as two copies in *A. nidulans*, while a single copy exists in the other species.

D-Galactose metabolism

D-Galactose is a common monosaccharide that can be utilized by all living organisms. Most microorganisms, including filamentous fungi, can utilize D-galactose for growth via three pathways: the Leloir pathway, the oxidoreductive pathway (also referred to as the De Ley-Doudoroff pathway) [73], and the non-phosphorylated De Ley-Doudoroff pathway (Figure 2).

The main route of fungal D-galactose metabolism is the Leloir pathway, resulting in its conversion to glucose 6-phosphate, which can enter glycolysis (Figure 2). In the first step of this pathway, β-D-galactose, which is the most common form released during plant polysaccharides degradation, is epimerized to α-D-galactose by galactose mutarotase (also known as aldose 1-epimerase, GalM, EC 5.1.3.3). Subsequent reactions catalyzed by galactokinase (GalK, EC 2.7.1.6), D-galactose-1-phosphate uridylyl transferase (GalT, EC 2.7.7.12) and UDP-galactose 4-epimerase (GalE, EC 5.1.3.2) convert D-galactose to D-glucose 1-phosphate, which is finally transformed to D-glucose 6-phosphate by the action of phosphoglucomutase (PgmB, EC 5.4.2.2). The Leloir pathway is highly conserved and we observed that all five enzymes involved in this pathway are present in the six studied species with at least one copy. The GalK and GalT encoding genes are extremely conserved, and present as a single copy in each fungus, with the exception of two copies of *galK* in *D. squalens*. In addition, almost all genes involved in the Leloir pathway of *A. nidulans* identified in this study have been previously experimentally characterized [74].

In the oxidoreductive pathway, β-D-galactose is converted into D-fructose by a series of reductive and oxidative steps via the intermediates galactitol, L-xylo-3-hexulose, D-sorbitol, and D-fructose [73]. The enzymes successively involved in this pathway include aldose reductase (EC 1.1.1.21), galactitol dehydrogenase (LadB, EC 1.1.1.-), L-xylo-3-hexulose reductase (XhrA, EC 1.1.1.-), and D-sorbitol dehydrogenase (GutB, EC 1.1.1.14). At the end, D-fructose is then phosphorylated by fructokinase (EC 2.7.1.4) to D-fructose

6-phosphate, which enters glycolysis. Previously studies in *A. niger* and *T. reesei* showed that two distinct enzymes were involved in catalyzing the first enzymatic step of the oxidoreductive pathway. In *T. reesei* the conversion of galactitol to L-xylo-3-hexulose was mediated by L-arabinitol dehydrogenase (LAD1) [75], while in *A. niger* a different gene, the D-galactitol dehydrogenase (LadB), instead of ortholog gene of LAD1, has been identified for catalyzing this reaction [76]. Similarly, the ortholog of LadB, AN4336 in *A. nidulans* was predicted for this conversion. However, different from other species the product of galactitol oxidation was suggested as L-sorbose in *A. nidulans* [77-80]. L-sorbose was produced from D-galactitol oxidation catalyzed by LadA, and it was then reduced to D-sorbitol by the L-xylulose reductase LxrA [77, 81]. In line with this hypothesis, LadA (AN0942) associated with the production of L-sorbose was suggested in a recent study [74]. In our study, we predicted the same gene set involved in this oxidoreductive pathway in *A. nidulans* as previously reported [79, 80, 82]. Although the oxidoreductive pathway for D-galactose catabolism was well established in *Aspergillus* species and *T. reesei*, knowledge of the oxidoreductive pathway in *Penicillium* and Basidiomycota species is scarce. As shown in Figure 2, the enzymes involved in reduction of the galactitol into L-xylo-3-hexulose were not identified in *P. subrubescens* and the two Basidiomycota species. Nevertheless, we identified the potential genes related to the other four reactions, indicating presence of a possible alternative enzymatic conversion between galactitol and L-xylo-3-hexulose in these species.

In the non-phosphorylated De Ley-Doudoroff pathway, D-galactose is eventually metabolized into pyruvate and D-glyceraldehyde-3-phosphate through five consecutive reactions producing six intermediates including D-galactono-1,4-lactone, D-galactonate, 2-dehydro-3-deoxy-D-galactonate, pyruvate, D-glyceraldehyde and D-glyceraldehyde-3-phosphate, which are catalyzed by five specific enzymes: D-galactose dehydrogenase (EC 1.1.1.48), gamma 1,4 lactonase (EC 3.1.1.25), D-galactonate dehydratase

(DGD, EC 4.2.1.6), an aldolase (EC 4.1.2.51) and a dihydroxyacetone kinase (DAK, EC 2.7.1.28). Pyruvate enters the TCA and glyoxylate cycles while D-glyceraldehyde-3-phosphate enters glycolysis. With the exception of the enzymes synthesizing the penultimate step in this pathway, which are still elusive, the corresponding genes involved in other reactions have been identified for all species in our study. Most species harbor two D-galactose dehydrogenase encoding genes but *T. reesei* has only one. The gamma 1,4 lactonase is highly conserved in all six tested species with one copy, while DGD encoding gene has multiple copies, especially in *P. subrubescens* and *T. reesei* that contain seven and eight copies, respectively. The DAK encoding gene was detected in all species with one copy except in *T. reesei* with two duplicates.



Overall, the D-galactose metabolism is relatively conserved in all studied species, despite several specific enzymes encoding genes that differ in copy numbers depending on the species and their taxonomic distances. Notably, the predicted gene content of D-galactose metabolism pathway in the two Basidiomycota species showed the most differences from the four Ascomycota species (Supplemental Table S3).

*3.3.3 Sugar catabolism with low conservation*

Pentose catabolic pathway (PCP)

L-Arabinose and D-xylose are the most abundant pentoses found in the hemicelluloses (such as (arabino)xylan and xyloglucan), and pectin of the plant cell wall [77]. Typically, L-arabinose and D-xylose are metabolized by the fungus through the pentose catabolic pathway (PCP), consisting of oxidation, reduction, and phosphorylation reactions to form D-xylulose-5-phosphate, which enters the PPP.

PCP is present in the majority of filamentous fungi, although there are some differences in the enzymes that catabolize each specific step. Traditionally, only a single enzyme was assigned to catalyze each reaction in PCP. A recent

study revealed the redundancy and complexity of the conversion of pentose sugars in *A. niger* [9]. L-arabinose goes through the sequential reactions: reduction catalyzed by NADPH-dependent L-arabinose reductase (LarA) and D-xylose reductase (XyrA/XyrB) (EC 1.1.1.21); oxidation catalyzed by L-arabitol-4-dehydrogenase (LadA, EC 1.1.1.12), xylitol dehydrogenase (XdhA), and sorbitol dehydrogenase (SdhA) (EC 1.1.1.12); reduction catalyzed by L-xylulose reductases (LxrA/LxrB, EC 1.1.1.10) [83]; and again oxidation (LadA/XdhA/SdhA, EC 1.1.1.9), by which L-arabinose is progressively converted to L-arabitol, L-xylulose, xylitol and D-xylulose (Figure 2). D-xylulose is further phosphorylated into D-xylulose 5-phosphate by D-xylulose kinase (XkiA, EC 2.7.1.17).

D-Xylose metabolism starts with the reduction of D-xylose into xylitol catalyzed by the same enzymes (LarA/XyrA/XyrB, EC 1.1.1.307) involved in the reduction of L-arabinose. Xylitol is further converted into D-xylulose and D-xylulose 5-phosphate sequentially as described in L-arabinose metabolism. Based on our prediction, the genes encoding the enzymes for the D-xylose catabolic pathways are present in all studied species, in some cases as multiple copies, particularly the D-xylose reductase and xylitol dehydrogenase genes (Table S3). In contrast, not all the genes involved in the metabolism of L-arabinose are conserved across our studied fungi, e.g., no orthologs of *A. niger* LadA and LxrA/LxrB genes were detected in the two Basidiomycota species. In addition, instead of two copies of the gene encoding LadA in *A. niger*, only a single LadA gene was found in the other two studied Eurotiomycetes (*A. nidulans* and *P. subrubescens*), and in one Sordariomycete species (*T. reesei*).

Given the resemblance between the reduction of L-arabinose and the first reduction reaction of the D-galactose oxidoreductive pathway, it is not a surprise that the same enzymes were reported for catalyzing the production of L-arabitol/xylitol and galactitol in *T. reesei* and *A. nidulans* [76, 77, 84]. In this study, we predicted similar functions of these enzyme-encoding genes for

the other four studied species (Supplemental Table S3).

Intriguingly, we predicted that the orthologs of *A. niger* LadA and LadB seem to have broader functions in other studied species than their predicted or characterized functions in *A. niger*. For example, we predicted that LAD1 (TrB1462W) of *T. reesei*, homologous to *A. niger* LadA, is involved in four enzymatic conversions: L-arabitol to L-xylulose, xylitol to D-xylulose, galactitol to L-xylo-3-hexulose and D-sorbitol to keto-D-fructose. The two reactions, L-arabitol to L-xylulose and galactitol to L-xylo-3-hexulose, were previously reported [84]. Exceptions are the conversions of galactitol to L-xylo-3-hexulose in *P. subrubescens* and the two Basidiomycota species, in which other genes seem to take over this function (Supplemental Table S3). Similarly, LadB (AN4336) in *A. nidulans* was predicted to be involved in multiple reactions (Supplemental Table S3).



L-Rhamnose metabolism

L-Rhamnose is enriched in some fractions of plant biomass, such as hemicellulose and pectin. In fungi, L-rhamnose is generally degraded in four consecutive reactions catalyzed by a NADH-dependent L-rhamnose-1-dehydrogenase (LraA, EC 1.1.1.173), L-rhamnonic acid lactonase (LrlA, EC 3.1.1.65), L-rhamnonate dehydratase (LrdA, EC 4.2.1.90) and 2-keto-3-deoxy-L-rhamnonate aldolase (LkaA, EC 4.1.2.53), resulting in the formation of L-rhamnose-1,4-lactone, L-rhamnonate, 2-dehydro-3-deoxy-l-rhamnonate, pyruvate and L-lactaldehyde [14, 85].

All genes involved in the four enzymatic steps of this pathway have been first identified in the yeast *Scheffersomyces (Pichia) stipitis* [86], and the *A. niger* L-rhamnose catabolic genes have also recently been characterized [14, 85]. Therefore, based on six known genes in *A. niger* (Table S3), we predicted the genes associated with this pathway in the other species. The genes encoding enzymes involved in the first and last (fourth) reactions were highly conserved in all six studied species, except the two copies of the genes encoding LraA and LkaA observed for *A. niger* and a single gene for other

species. However, the LrlA and LrdA encoding genes involved in the second and third reactions were only conserved in three studied Ascomycetes, but missing in the two studied Basidiomycetes (Table S3), which may indicate presence of possible alternative enzymes involved in these reactions in the Basidiomycetes. In addition, we identified the same enzyme candidates of L-rhamnose metabolism in *A. nidulans* as in a previous study [87], with the exception of LkaA.

D-Gluconate metabolism

A wide group of filamentous fungi have the ability to produce gluconic acid, such as *Aspergillus* and *Penicillium* [88]. In *A. niger*, D-gluconate catabolism includes five enzymatic reactions which are catalyzed by five enzymes: glucose oxidase (GOX, EC 1.1.3.4), D-glucono-1,5-lactone lactonodehydrolase or gluconolactonase (EC 3.1.1.17), gluconate dehydratase (EC 4.2.1.140), 2-dehydro-3-deoxy-D-gluconate D-glyceraldehyde-lyase (EC 4.1.2.51) and gluconate kinase (GukA, EC 2.7.1.12).

In all the studied species, missing enzymes in the D-gluconate catabolism were observed (Figure 2). Furthermore, the gene(s) involved in the conversion of 2-keto-3-deoxy-D-gluconate to pyruvate remains unknown among all the studied fungi. Apart from *P. subrubescens*, the other fungi have one or more copies of glucose oxidase (GOX). The highest number of GOX encoding genes was identified in *P. chrysosporium* with five genes, followed by *D. squalens* with three genes. The genes encoding gluconolactonase (EC 3.1.1.17) involved in the second step were present with more than two copies in most species except *P. chrysosporium*. The gluconate dehydratase (EC 4.2.1.140) is highly conserved in all species. In addition, D-gluconate can enter the PPP through the catalysis by GukA, which was identified in most of the studied fungi, except *D. squalens*.

### *3.4 The expression profiling of sugar metabolic genes in each fungus*

In order to explore transcriptional responses of different fungi to distinct

monosaccharides, we performed RNA-seq experiments from fungal cultivations on nine monosaccharides. The clustering results indicate the similarity of overall expression profile across different growth conditions. The hexose (D-glucose, D-mannose and D-fructose) and pentose (D-xylose and L-arabinose) conditions were well clustered to each other, respectively, on the transcriptome data of the three Eurotiomycetes (*A. niger*, *A. nidulans*, *P. subrubescens*) (Figure 3 and Figure S1), while in the other species L-arabinose was clustered more closely with D-galactose than with D-xylose. In addition, the samples of D-galacturonic acid and D-glucuronic acid were also closely clustered together for all studied species.



Furthermore, we observed obvious sugar-specific inducing patterns for many SMGs during fungal growth on the corresponding sugars (Figure 3 and Figure S1). These clear inducing patterns of specific sugar metabolic genes on corresponding growth conditions support their predicted function in our sugar metabolic models. For example, the predicted galacturonic acid metabolism related genes were significantly more highly expressed during fungal growth on galacturonic acid than their expression on other sugar conditions for all tested fungi, except for *D. squalens* that lacks the corresponding transcriptome data. In addition, many of the L-rhamnose and PCP-related genes were induced on the corresponding growth conditions in *A. niger*, *A. nidulans*, *P. subrubescens* and *T. reesei*. A few D-galactose oxidoreductive pathway related genes were also induced for the three tested Eurotiomycetes, while no clear inducing patterns were observed for the other tested species. Several genes that were involved in both PCP and the D-galactose oxidoreductive pathway, were highly expressed on D-xylose, L-arabinose and D-galactose in *A. nidualans*, *P. subrubescens* and *T. reesei*. In contrast to the evident sugar-specific inducing patterns of SMGs observed for Ascomycetes, only a limited number of SMGs induced by a corresponding sugar were observed for the two studied Basidiomycetes.

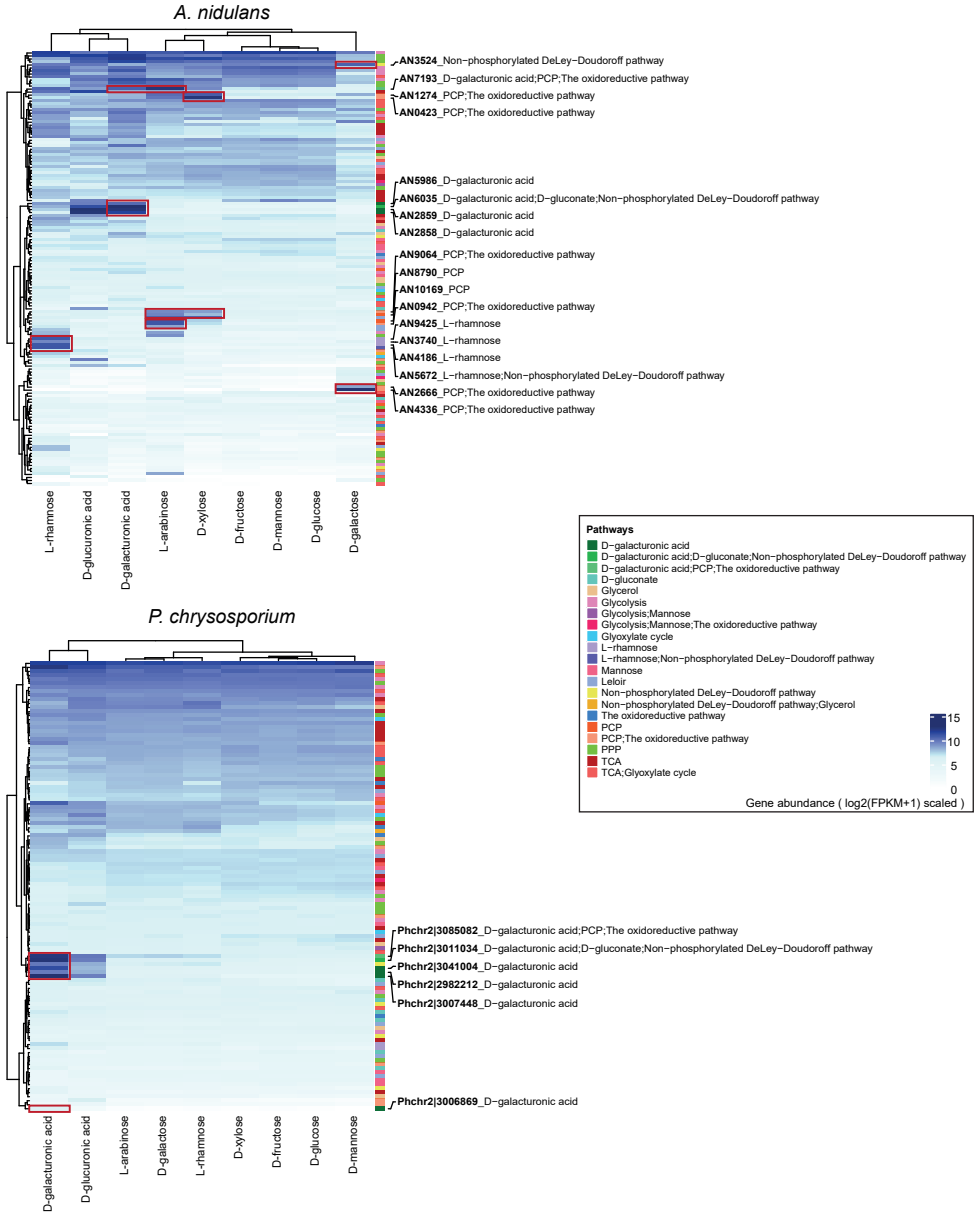


**Figure 3. Expression profiles of sugar metabolism-related genes in *A. nidulans* and *P. chrysosporium* during their growth on diverse monosaccharides.** The blue color from light to dark indicates a gene expression level from low to high. Selected genes with specific sugar induced expression patterns are marked with a red box on the heatmap, and their gene IDs and the associated pathways are displayed. On the right bar, different colors indicate different pathways.

### *3.5 Proteome profile of sugar metabolism-related genes (SMGs) showed similar sugar inducing pattern as transcriptome data*

The intracellular proteome of the studied fungi (except *D. squalens*) grown on each monosaccharide condition was also analyzed to reveal the fungal response to each specific sugar at the proteome level. In general, the protein abundance profile of SMGs showed higher similarity among replicates of the same sugar, as well as more stable production among different sugars compared to the corresponding production profiles analyzed using the whole proteome (Figure S2).

The comparison of protein abundance profile of each specific sugar pathway among different sugars showed similar patterns as observed in the transcriptome data. The abundance profiles of proteins involved in sugar pathways of the D-galacturonic acid, L-rhamnose and PCP in the studied Ascomycetes (*A. niger*, *A. nidulans*, *P. subrubescens* and *T. reesei)* match their predicted functions. As observed, the protein abundances of SMGs involved in the above pathways were higher in corresponding sugars than in other tested sugars (Figure 4A and Figure S3). In contrast, only proteins involved in the D-galacturonic acid pathway showed strong sugar specificity in *P. chrysosporium*, while the sugar inducing patterns of other SMGs were not obvious in the proteome data (Figure 4B). For other studied sugar pathways (e.g. TCA, glycolysis and PPP), we did not observe clear difference in protein abundance among sugars.

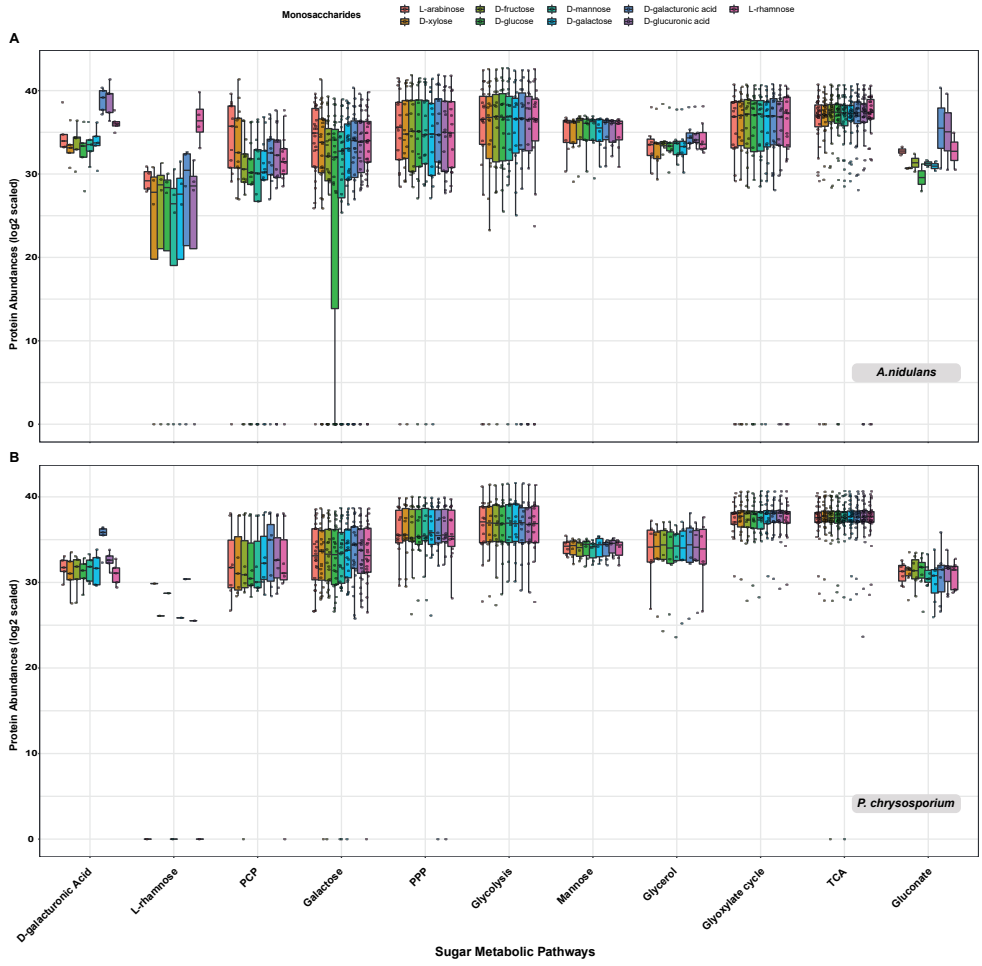


**Figure 4. Protein abundance profiles of the sugar metabolism-related proteins involved in each sugar metabolic pathway in *A. nidulans* (A) and *P. chrysosporium* (B).** On the boxplot, different colors indicate different monosaccharide growth conditions, and each small circle indicates an individual protein related to each specific sugar metabolic pathway. The metabolic enzymes that were detected in less than one third of the tested growth conditions were not included in the boxplots. The y-axis represents the abundance of proteins (log2 scaled), and the x-axis depicts different sugar metabolic pathways.

## *3.6 Correlation between the abundance of metabolites and gene expression levels of sugar metabolism-related genes (SMGs)*

To further verify the accuracy of our predicted sugar metabolic pathways and

investigate the association between metabolites and SMGs, we compared the metabolomics profiles across different sugar conditions and performed the correlation analysis between the abundance of SMGs and metabolites in each fungus (except *D. squalens* that lacks related data). Overall, we detected the sugars taken up by the cell, sugar metabolic intermediates and several storage compounds for each fungus (Figure S4). Several intermediates were consistently detected with higher abundance in corresponding sugar condition for most of studied fungi, such as xylitol, galactitol and galactonic acid showed higher abundance in D-xylose, D-galactose and D-galacturonic acid conditions for most of studied fungi. The commonly detected storage compounds include the erythritol, mannitol and ribitol.



Consistent with stronger inducing patterns of SMGs identified in the transcriptome and proteome profiles of Ascomycetes compared to Basidiomycete, we detected that more SMGs showed stronger correlation with the corresponding metabolites in four studied Ascomycetes than in *P. chrysosporium* (Figure 5 and Figure S5). For instance, in total 21 SMGs in *A. nidulans* (mainly including PCP, D-galactose, D-galacturonic acid, L-rhamnose pathways and glycolysis) showed strong positive correlations (correlation coefficient $\geq 0.8$) with their corresponding metabolites (Figure 5). However, only two SMGs displayed high positive correlations with the metabolites in *P. chrysosporium*. In the other two studied Eurotiomycetes species, we identified 14, 1, 2, 1 and 1 SMGs, respectively, involved in the PCP, D-galacturonic acid, L-rhamnose, glycolysis and glycerol pathways in *A. niger*, as well as 3, 3, 2 and 4 SMGs respectively involved in the PCP, D-galacturonic acid, L-rhamnose, and D-galactose pathways in *P. subrubescens,* which strongly correlated to detection of the related metabolites (Figure S5). As for *T. reesei*, we discovered six PCP-related SMGs, which have high correlations with xylitol (1), D-arabinose (3), and D-xylose (2), two glycolysis-related SMGs highly associated with D-glucose (1) and D-glucose-6-phosphate (1), but no SMGs showed high correlations with rhamnose (Figure S5). Taken together,

the abundance profile of SMGs and metabolites confirmed the predicted sugar metabolic pathways and the diversity of fungal responses to different sugars.


**Figure 5. Correlation between the abundance of metabolites and sugar metabolism-related genes in *A. nidulans* and *P. chrysosporium*.** The y-axis represents Pearson correlation coefficient (PCC ≥ 0.5), and x-axis depicts different metabolites and the related sugar metabolic pathways (highlighted in blue). Each dot represents a gene and its color indicates the corresponding sugar pathway that it is involved in. The positive and negative correlation were shown in filled and open circles, respectively. Only the names of genes with

high correlation (PCC ≥ 0.8) to the analyzed metabolites are displayed.

### *3.7 Comparative growth profiles on different carbon sources*

We performed growth profiling of the six fungi on nine different monosaccharides as solo carbon sources to analyze the influence of different sugars on their growth. Given that the completeness and efficiency of sugar metabolic pathways are directly linked with their growth on the corresponding sugar, we expect clear differences in growth profiles to be observed when two different fungi with predicted different completeness of specific sugar metabolic pathways are grown on the same sugar. Therefore, the comparison of growth profiles of these fungi on different sugars could partially confirm the accuracy of our predicted sugar metabolic pathways.



The growth profiles of the six fungi showed both notable similarities and differences on different sugars (Figure 6). In line with the complete sugar metabolic pathways of all tested sugars predicted for *A. niger* and the other two studied Eurotiomycetes species (*A. nidulans* and *P. subrubescens*), they showed similar growth profiles on all tested carbon sources. In addition, we noticed good growth for all six species on D-glucose, D-fructose and D-mannose compared to their growth on no-carbon source, which matches the conservation and completeness of the sugar metabolic pathways of these hexoses across all studied species. In contrast, different species show varying degrees of growth on L-rhamnose. *A. nidulans* and *P. subrubescens* grew better on this monosaccharide than *T. reesei*, while the two Basidiomycota species (*P. chrysosporium* and *D. squalens*) grew poorly on L-rhamnose, which can be explained by the missing enzymes of L-rhamnose metabolic pathway predicted for these species in our study. In addition, there is a clear difference in growth on L-arabinose between the Ascomycota and Basidiomycota species. The poor growth on L-arabinose for *P. chrysosporium* and *D. squalens* could be linked to the predicted absence of both *ladA* and *lxrA*/*lxrB* encoding genes in their genomes. Although orthologues of *xdhA* and *sdhA* were detected in the two Basidiomycete species, they cannot fully compensate for the loss of

*ladA,* which seems to play a key role in L-arabinose metabolism. In particular, growth on L-arabinose was almost abolished for *P. chrysosporium*.

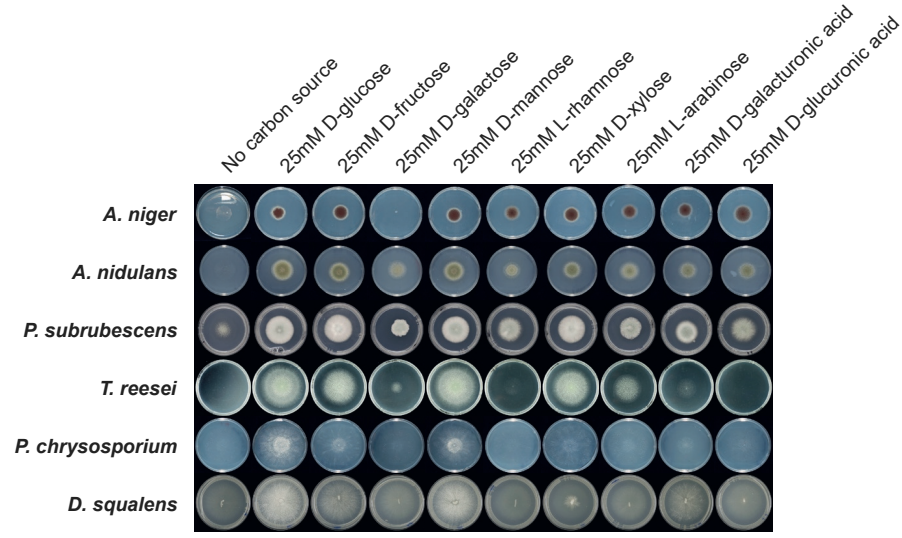


**Figure 6. Fungal growth profiling on different monosaccharides.** All strains were grown on MM with nine different carbon sources. Growth was performed at 30°C for the two Aspergilli and 25°C for other species. If growth on a specific carbon source is the same as with no carbon source, it is considered no growth.

Besides the consistency observed between growth profile and predicted sugar pathway models, we also observed the scenario that despite the fact that fungi possess a full enzyme set in a sugar pathway, it does not ensure their good growth on the related sugar. For example, although the genes related to D-xylose metabolism were present in all fungi, the two Basidiomycota species showed poor growth on D-xylose. A similar profile was observed for the poor growth of *T. reesei* on D-galacturonic acid although the full enzyme set to metabolize the D-galacturonic acid is present in its genome. In *A. niger* and *A. nidulans*, all genes involved in D-ribose metabolism were predicted, but their growth on D-ribose was poor. However, *P. subrubescens* and *T. reesei* grew well on D-ribose that is in line with the presence of the related genes in their genome. Due to the absence of ribokinase encoding genes in *P. chrysosporium* and *D. squalens*, poor growth on D-ribose was observed. A strong difference

in growth on D-galactose was observed between *A. niger*, the other two Eurotiomycetes species, and the two Basidiomycota species. Although *A. niger* and *T. reesei* have a full set of D-galactose metabolic enzymes, the growth of *A. niger* and *T. reesei* on D-galactose was poor. *P. chrysosporium* and *D. squalens* also grew poorly on D-galactose, while *A. nidulans* and *P. subrubescens* displayed relatively good growth on D-galactose.

## 4. Discussion

We successfully established sugar metabolic networks among six diverse fungi using an orthology-based approach and integrating of a broad range of functional annotation and transcriptome data. The difference of predicted sugar metabolic pathways between the reference species (*A. niger*) and the other five studied fungi is well correlated with their taxonomic distances, e.g., most sugar catabolic pathways were highly conserved between the other two studied Eurotiomycetes and *A. niger*, while a larger level of diversity was observed between the Sordariomycete species and *A. niger*, and even more so for the Basidiomycete species. The predicted sugar metabolic pathways and their corresponding genes were partially supported by the growth profiles and multi-omics profiles of fungi grown on diverse sugars.

Many of our predicted SMGs in the two well studied model fungi (*A. nidulans* and *T. reesei*) have already been experimentally characterized (Supplemental Table S3), which confirmed the accuracy of our prediction. For the less studied *P. subrubescens* and the two Basidiomycete species, we provided the first sugar metabolic blueprint of these species.

Compared to the fungal sugar metabolic pathway information available in the well-known The Kyoto Encyclopedia of Genes and Genomes (KEGG) database (https://www.kegg.jp/), we provided a more comprehensive and detailed sugar metabolic network. For instance, there is no specific pathway assigned to D-galacturonic acid metabolism in the KEGG database, and galactose metabolic pathway (KEGG pathway: map00052) was not explicitly

delineated into three different sub-pathways. Some crucial reactions of PCP and PPP pathways (KEGG pathway: map00030 and map00040, respectively) were not well-annotated. For example, four reactions (EC, 2. 7.1.47/1.1.3.4/4.2.1.140/4.1.2.51) involved in PPP and three reactions (EC, 1.1.1.21/1.1.1.12/1.1.1.10) in PCP were absent in KEGG for *A. nidulans* and *T. reesei*.

Several sugar metabolic pathways were highly conserved across all studied fungi, e.g., TCA, glycolysis, PPP and D-galacturonic acid metabolism. Other pathways showed large variation across the different fungi, including L-rhamnose and L-arabinose catabolism. However, the completeness of the predicted pathways cannot ensure a very similar transcriptome response and growth profile of the fungi grown on corresponding sugars. For instance, although the completeness of pathways was identified for D-fructose and D-galacturonic acid in all studied fungi, we observed the relatively poor growth of *P. chrysosporium* on D-fructose, poor growth of *T. reesei* and *P. chrysosporium* on D-galacturonic acid. This could be related to the difference of catalytic efficiency and regulation mechanisms of corresponding metabolic enzymes in different species. In addition, the efficiency of fungal sugar utilization is not only related to metabolic enzymes, but also to the sugar transport system [89]. One of the best examples to support this hypothesis is that the poor growth of *A. niger* on D-galactose was mainly due to the inactivity of relevant transport and germination triggers [90-92]. On the other hand, the predicted incompleteness of a metabolic pathway for a sugar does not always mean a poor growth profile. For example, we failed to identify one key enzyme converting L-xylulose to xylitol in L-arabinose catabolism in the two Basidiomycota species, however, the corresponding intermediates (e.g. arabitol, xylitol and D-xylulose) were detected in metabolomics analysis of *P. chrysosporium* and the growth profile showed that they can utilize L-arabinose. This indicates that possible new enzymes may be involved in this pathway. The failure of identification of these alternative enzymes in our

study could be due to their low sequence similarity to the enzymes predicted or validated in *A. niger*. An alternative approach integrating sequence features and multiple omics data using machine learning could be used to detect those enzymes [93-95].

The different transcriptome and proteome profile of SMGs between Ascomycota and Basidiomycota species could be related to their different transcriptional regulation and be linked with their adaption to different ecological niches. In line with this, previous studies have demonstrated a dramatic difference in the repertoire of regulatory proteins of Ascomycete and Basidiomycete fungi [96, 97]. Orthologs of the well-studied transcription factors controlling specific SMGs in Ascomycete fungi, e.g., XlnR, AraR, RhaR, GaaR, and GalX, were missing in the Basidiomycete fungi. The lack of clear sugar inducing patterns in the Basidiomycota species could also suggest that in their natural environment, the studied Basidiomycetes rarely experience the relatively high monomeric-sugar concentrations that were used in our study.



## Author Contributions

All authors contributed to the manuscript. R.P.d.V. conceived and supervised the overall project. J.L. performed the formal analysis and wrote the original draft. T.C., S.G., R.S.K., J.M., and S.S.-C. performed the cultivations and sample preparation. MVA-P analyzed A. niger pathway. Y.Z., S.T., A.L., V.N. and I.V.G. performed and coordinated the RNAseq experiments and initial analyses. C.S.C. and N.T. performed the proteomics and metabolomics analysis. A.T., M.R.M. and B.S. were involved in the analysis and discussions. M.P. and R.P.d.V. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

## Funding

J.L. was supported by the China Scholarship Council (CSC 201909110079). T.C. was supported by a grant of the Dutch Research Council (NWO)

ALWOP.233 granted to R.P.d.V. S.G., and R.S.K. were supported by a grant of the Applied Science Division (TTW) of the Dutch Research Council (NWO) and the Biotechnology and Safety Program of the Ministry of Infrastructure and Water Management 15807, and R.P.d.V. and J.M. were supported by the Chinese Scholarship Council (CSC 201907720027). S.S.C. was supported by a Postdoctoral Researcher fellowship from Mexico Government (CONACYT, support 360912). The Academy of Finland grants numbers 308284 and 348443 to M.R.M. are acknowledged.

## Acknowledgements

This research was performed on a project award (proposal: 10.46936/fics.proj.2018.50379/60006403) under the FICUS program and used resources at the DOE Joint Genome Institute and the Environmental Molecular Sciences Laboratory, which are DOE Office of Science User Facilities. Both facilities are sponsored by the Biological and Environmental Research program and operated under Contract Nos. DE-AC02-05CH11231 (JGI) and DE-AC05-76RL01830 (EMSL).

## Conflicts of Interest

The authors declare no conflicts of interest.

## References

1. Guerriero, G., et al., *Lignocellulosic biomass: biosynthesis, degradation, and industrial utilization.* **Engineering in Life Sciences**, 2016. 16(1): p. 1-16.
2. de Vries, R.P. and J. Visser, *Aspergillus enzymes involved in degradation of plant cell wall polysaccharides.* **Microbiology and Molecular Biology Reviews**, 2001. 65(4): p. 497-522.
3. Culleton, H., V. McKie, and R.P. de Vries, *Physiological and molecular aspects of degradation of plant polysaccharides by fungi: what have we learned from Aspergillus?* **Biotechnology Journal**, 2013. 8(8): p. 884-894.
4. de Vries, R.P. and M.R. Mäkelä, *Genomic and postgenomic diversity of fungal plant biomass degradation approaches.* **Trends in Microbiology**, 2020. 28(6): p. 487-499.
5. Khosravi, C., et al., *Sugar catabolism in Aspergillus and other fungi related to the*

*utilization of plant biomass.* **Advances in Applied Microbiology**, 2015. 90: p. 1-28.
6. Mäkelä, M.R., N. Donofrio, and R.P. de Vries, *Plant biomass degradation by fungi.* **Fungal Genetics and Biology**, 2014. 72: p. 2-9.
7. Panneman, H., et al., *Cloning and biochemical characterisation of an Aspergillus niger glucokinase: Evidence for the presence of separate glucokinase and hexokinase enzymes.* **European Journal of Biochemistry**, 1996. 240(3): p. 518-525.
8. Panneman, H., et al., *Cloning and biochemical characterisation of Aspergillus niger hexokinase: The enzyme is strongly inhibited by physiological concentrations of trehalose 6-phosphate.* **European Journal of Biochemistry**, 1998. 258(1): p. 223-232.
9. Chroumpi, T., et al., *Revisiting a 'simple' fungal metabolic pathway reveals redundancy, complexity and diversity.* **Microbial Biotechnology**, 2021. 14: p. 2525-2537.
10. Aguilar-Pontes, M.V., et al., *The gold-standard genome of Aspergillus niger NRRL 3 enables a detailed view of the diversity of sugar catabolism in fungi.* **Studies in Mycology**, 2018. 91: p. 61-78.
11. Kubicek, C.P., *Regulatory aspects of the tricarboxylic acid cycle in filamentous fungi - a review.* **Transactions of the British Mycological Society**, 1988. 90(3): p. 339-349.
12. Strijbis, K. and B. Distel, *Intracellular acetyl unit transport in fungal carbon metabolism.* **Eukaryotic Cell**, 2010. 9(12): p. 1809-1815.
13. Kunze, M., et al., *A central role for the peroxisomal membrane in glyoxylate cycle function.* **Biochimica et Biophysica Acta (BBA) - Molecular Cell Research**, 2006. 1763(12): p. 1441-1452.
14. Chroumpi, T., et al., *Identification of a gene encoding the last step of the L-rhamnose catabolic pathway in Aspergillus niger revealed the inducer of the pathway regulator.* **Microbiological Research**, 2020. 234: p. 126426.
15. Li, L., C.J. Stoeckert, and D.S. Roos, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* **Genome Research**, 2003. 13(9): p. 2178-2189.
16. Emms, D.M. and S. Kelly, *OrthoFinder: phylogenetic orthology inference for comparative genomics.* **Genome Biology**, 2019. 20(1): p. 1-14.
17. Grigoriev, I.V., et al., *MycoCosm portal: gearing up for 1000 fungal genomes.* **Nucleic Acids Research**, 2014. 42(D1): p. D699-D704.
18. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using DIAMOND.* **Nature Methods**, 2015. 12(1): p. 59-60.
19. Kelly, S. and P.K. Maini, *DendroBLAST: approximate phylogenetic trees in the absence of multiple sequence alignments.* **PLoS One**, 2013. 8(3): p. e58537.
20. Karp, P.D., S. Paley, and P. Romero, *The pathway tools software.* **Bioinformatics**, 2002. 18(suppl_1): p. S225-S232.

21. Karp, P.D., et al., *Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology.* **Briefings in Bioinformatics**, 2021. 22(1): p. 109-126.
22. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* **Nucleic Acids Research**, 2000. 28(1): p. 27-30.
23. Consortium, G.O., *The gene ontology project in 2008.* **Nucleic Acids Research**, 2008. 36(suppl_1): p. D440-D444.
24. Hunter, S., et al., *InterPro: the integrative protein signature database.* **Nucleic Acids Research**, 2009. 37(suppl_1): p. D211-D215.
25. Petersen, T.N., et al., *SignalP 4.0: discriminating signal peptides from transmembrane regions.* **Nature Methods**, 2011. 8(10): p. 785-786.
26. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* **BMC Bioinformatics**, 2003. 4(1): p. 1-14.
27. Sonnhammer, E.L., S.R. Eddy, and R. Durbin, *Pfam: a comprehensive database of protein domain families based on seed alignments.* **Proteins: Structure, Function, and Bioinformatics**, 1997. 28(3): p. 405-420.
28. Mistry, J., et al., *Pfam: The protein families database in 2021.* **Nucleic Acids Research**, 2021. 49(D1): p. D412-D419.
29. Bairoch, A. and R. Apweiler, *The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.* **Nucleic Acids Research**, 2000. 28(1): p. 45-48.
30. Keseler, I.M., et al., *EcoCyc: fusing model organism databases with systems biology.* **Nucleic Acids Research**, 2013. 41(D1): p. D605-D612.
31. Caspi, R., et al., *The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* **Nucleic Acids Research**, 2014. 42(D1): p. D459-D471.
32. Gruben, B.S., et al., *Expression-based clustering of CAZyme-encoding genes of Aspergillus niger.* **BMC Genomics**, 2017. 18(1): p. 1-18.
33. Casado López, S., et al., *Induction of genes encoding plant cell wall-degrading carbohydrate-active enzymes by lignocellulose-derived monosaccharides and cellobiose in the white-rot fungus Dichomitus squalens.* **Applied and Environmental Microbiology**, 2018. 84(11): p. e00403-18.
34. de Vries, R.P., et al., *A new black Aspergillus species, A. vadensis, is a promising host for homologous and heterologous protein production.* **Applied and Environmental Microbiology**, 2004. 70(7): p. 3954-3959.
35. Klaubauf, S., et al., *Similar is not the same: differences in the function of the (hemi-) cellulolytic regulator XlnR (Xlr1/Xyr1) in filamentous fungi.* **Fungal Genetics and Biology**, 2014. 72: p. 73-81.
36. Eastwood, D.C., et al., *The plant cell wall–decomposing machinery underlies the functional diversity of forest fungi.* **Science**, 2011. 333(6043): p. 762-765.

37. Daly, P., et al., *Colonies of the fungus Aspergillus niger are highly differentiated to adapt to local carbon source variation.* **Environmental Microbiology**, 2020. 22(3): p. 1154-1166.
38. Kim, S., N. Gupta, and P.A. Pevzner, *Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases.* **Journal of Proteome Research**, 2008. 7(8): p. 3354-3363.
39. Kim, S. and P.A. Pevzner, *MS-GF+ makes progress towards a universal database search tool for proteomics.* **Nature Communications**, 2014. 5(1): p. 1-10.
40. Monroe, M.E., et al., *MASIC: a software program for fast quantitation and flexible visualization of chromatographic profiles from detected LC–MS (/MS) features.* **Computational Biology and Chemistry**, 2008. 32(3): p. 215-217.
41. Polpitiya, A.D., et al., *DAnTE: a statistical tool for quantitative analysis of-omics data.* **Bioinformatics**, 2008. 24(13): p. 1556-1558.
42. Chroumpi, T., et al., *Re-routing of sugar catabolism provides a better insight into fungal flexibility in using plant biomass-derived monomers as substrates.* **Frontiers in Bioengineering and Biotechnology**, 2021. 9: p. 167.
43. Hiller, K., et al., *MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis.* **Analytical Chemistry**, 2009. 81(9): p. 3429-3439.
44. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* **BMC Bioinformatics**, 2008. 9(1): p. 1-13.
45. Romano, A. and T. Conway, *Evolution of carbohydrate metabolic pathways.* **Research in Microbiology**, 1996. 147(6-7): p. 448-455.
46. Ruijter, G.J. and J. Visser, *Characterization of Aspergillus niger phosphoglucose isomerase. Use for quantitative determination of erythrose 4-phosphate.* **Biochimie**, 1999. 81(3): p. 267-272.
47. Ruijter, G., H. Panneman, and J. Visser, *Overexpression of phosphofructokinase and pyruvate kinase in citric acid-producing Aspergillus niger.* **Biochimica et Biophysica Acta (BBA) - General Subjects**, 1997. 1334(2-3): p. 317-326.
48. Pontremoli, S., et al., *Fructose diphosphatase from rabbit liver I. Purification and properties.* **Journal of Biological Chemistry**, 1965. 240(9): p. 3459-3463.
49. Jagannathan, V., K. Singh, and M. Damodaran, *Carbohydrate metabolism in citric acid fermentation. 4. Purification and properties of aldolase from Aspergillus niger.* **Biochemical Journal**, 1956. 63(1): p. 94.
50. Punt, P.J., et al., *Isolation and characterization of the glyceraldehyde-3-phosphate dehydrogenase gene of Aspergillus nidulans.* **Gene**, 1988. 69(1): p. 49-57.
51. Clements, J.M. and C.F. Roberts, *Molecular cloning of the 3-phosphoglycerate kinase (PGK) gene from Aspergillus nidulans.* **Current Genetics**, 1985. 9(4): p. 293-298.

52. Machida, M., et al., *Molecular cloning of a cDNA encoding enolase from the filamentous fungus, Aspergillus oryzae.* **Current Genetics**, 1996. 30(5): p. 423-431.
53. de Graaff, L., H. van den Broeck, and J. Visser, *Isolation and characterization of the Aspergillus niger pyruvate kinase gene.* **Current Genetics**, 1992. 22(1): p. 21-27.
54. Khitan, Z. and D.H. Kim, *Fructose: a key factor in the development of metabolic syndrome and hypertension.* **Journal of Nutrition and Metabolism**, 2013. 2013.
55. Jun, H., H. Guangye, and C. Daiwen, *Insights into enzyme secretion by filamentous fungi: comparative proteome analysis of Trichoderma reesei grown on different carbon sources.* **Journal of Proteomics**, 2013. 89: p. 191-201.
56. Stappler, E., et al., *Analysis of light-and carbon-specific transcriptomes implicates a class of G-protein-coupled receptors in cellulose sensing.* **Msphere**, 2017. 2(3): p. e00089-17.
57. Connerton, I., et al., *Comparison and cross-species expression of the acetyl-CoA synthetase genes of the ascomycete fungi, Aspergillus nidulans and Neurospora crassa.* **Molecular Microbiology**, 1990. 4(3): p. 451-460.
58. Ruijter, G.J., P.J. van de Vondervoort, and J. Visser, *Oxalic acid production by Aspergillus niger: an oxalate-non-producing mutant produces citric acid at pH 5 and in the presence of manganese.* **Microbiology**, 1999. 145(9): p. 2569-2576.
59. Richard, P. and S. Hilditch, *D-galacturonic acid catabolism in microorganisms and its biotechnological relevance.* **Applied Microbiology and Biotechnology**, 2009. 82(4): p. 597-604.
60. Martens-Uzunova, E.S. and P.J. Schaap, *An evolutionary conserved D-galacturonic acid metabolic pathway operates across filamentous fungi capable of pectin degradation.* **Fungal Genetics and Biology**, 2008. 45(11): p. 1449-1457.
61. Hilditch, S., *Identification of the fungal catabolic D-galacturonate pathway*. 2010, Espoo: VTT Technical Research Centre of Finland.
62. Kuorelahti, S., et al., *Identification in the mold Hypocrea jecorina of the first fungal D-galacturonic acid reductase.* **Biochemistry**, 2005. 44(33): p. 11234-11240.
63. Kuorelahti, S., et al., *L-galactonate dehydratase is part of the fungal path for D-galacturonic acid catabolism.* **Molecular Microbiology**, 2006. 61(4): p. 1060-1068.
64. Zhang, L., H. Thiewes, and J.A. van Kan, *The D-galacturonic acid catabolic pathway in Botrytis cinerea.* **Fungal Genetics and Biology**, 2011. 48(10): p. 990-997.
65. Alazi, E., et al., *The pathway intermediate 2-keto-3-deoxy-L-galactonate mediates the induction of genes involved in D-galacturonic acid utilization in Aspergillus niger.* **FEBS Letters**, 2017. 591(10): p. 1408-1418.
66. Klein, M., et al., *Glycerol metabolism and transport in yeast and fungi: established knowledge and ambiguities.* **Environmental Microbiology**, 2017. 19(3): p. 878-

893.

67. Nicol, R., K. Marchand, and W. Lubitz, *Bioconversion of crude glycerol by fungi.* **Applied Microbiology and Biotechnology**, 2012. 93(5): p. 1865-1875.
68. Hondmann, D.H., et al., *Glycerol catabolism in Aspergillus nidulans.* **Microbiology**, 1991. 137(3): p. 629-636.
69. Liepins, J., et al., *Enzymes for the NADPH-dependent reduction of dihydroxyacetone and D-glyceraldehyde and L-glyceraldehyde in the mould Hypocrea jecorina.* **The FEBS Journal**, 2006. 273(18): p. 4229-4235.
70. de Vries, R.P., et al., *Glycerol dehydrogenase, encoded by gldB is essential for osmotolerance in Aspergillus nidulans.* **Molecular Microbiology**, 2003. 49(1): p. 131-141.
71. Stryer, L., *Biochemistry*. 1995, New York: WH Freeman.
72. Kruger, N.J. and A. von Schaewen, *The oxidative pentose phosphate pathway: structure and organisation.* **Current Opinion in Plant Biology**, 2003. 6(3): p. 236-246.
73. Mojzita, D., et al., *L-xylo-3-hexulose reductase is the missing link in the oxidoreductive pathway for D-galactose catabolism in filamentous fungi.* **Journal of Biological Chemistry**, 2012. 287(31): p. 26010-26018.
74. Németh, Z., et al., *L-Arabinose induces D-galactose catabolism via the Leloir pathway in Aspergillus nidulans.* **Fungal Genetics and Biology**, 2019. 123: p. 53-59.
75. Pail, M., et al., *The metabolic role and evolution of L-arabinitol 4-dehydrogenase of Hypocrea jecorina.* **European Journal of Biochemistry**, 2004. 271(10): p. 1864-1872.
76. Mojzita, D., et al., *Identification of the galactitol dehydrogenase, LadB, that is part of the oxido-reductive D-galactose catabolic pathway in Aspergillus niger.* **Fungal Genetics and Biology**, 2012. 49(2): p. 152-159.
77. Seiboth, B. and B. Metz, *Fungal arabinan and L-arabinose metabolism.* **Applied Microbiology and Biotechnology**, 2011. 89(6): p. 1665-1673.
78. Fekete, E., et al., *The alternative D-galactose degrading pathway of Aspergillus nidulans proceeds via L-sorbose.* **Archives of Microbiology**, 2004. 181(1): p. 35-44.
79. Orosz, A., et al., *Metabolism of D-galactose is dispensable for the induction of the beta-galactosidase (bgaD) and lactose permease (lacpA) genes in Aspergillus nidulans.* **FEMS Microbiology Letters**, 2014. 359(1): p. 19-25.
80. Kowalczyk, J.E., et al., *Genetic interaction of Aspergillus nidulans galR, xlnR and araR in regulating D-galactose and L-arabinose release and catabolism gene expression.* **PLoS One**, 2015. 10(11): p. e0143200.
81. Flipphi, M., et al., *Biodiversity and evolution of primary carbon metabolism in*

*Aspergillus nidulans and other Aspergillus spp.* **Fungal Genetics and Biology**, 2009. 46(1): p. S19-S44.

82. Kulcsár, L., et al., *Identification of a mutarotase gene involved in D-galactose utilization in Aspergillus nidulans.* **FEMS Microbiology Letters**, 2017. 364(20).
83. Mojzita, D., et al., *The 'true'L-xylulose reductase of filamentous fungi identified in Aspergillus niger.* **FEBS Letters**, 2010. 584(16): p. 3540-3544.
84. Seiboth, B., et al., *The D-xylose reductase of Hypocrea jecorina is the major aldose reductase in pentose and D-galactose catabolism and necessary for β-galactosidase and cellulase induction by lactose.* **Molecular Microbiology**, 2007. 66(4): p. 890-900.
85. Khosravi, C., et al., *In vivo functional analysis of L-rhamnose metabolic pathway in Aspergillus niger: a tool to identify the potential inducer of RhaR.* **BMC Microbiology**, 2017. 17(1): p. 1-12.
86. Koivistoinen, O.M., et al., *Characterisation of the gene cluster for L-rhamnose catabolism in the yeast Scheffersomyces (Pichia) stipitis.* **Gene**, 2012. 492(1): p. 177-185.
87. MacCabe, A.P., et al., *Catabolism of L-rhamnose in A. nidulans proceeds via the non-phosphorylated pathway and is glucose repressed by a CreA-independent mechanism.* **Microbial Cell Factories**, 2020. 19(1): p. 1-15.
88. Shindia, A., et al., *Production of gluconic acid by some local fungi.* **Mycobiology**, 2006. 34(1): p. 22-29.
89. Peng, M., et al., *In silico analysis of putative sugar transporter genes in Aspergillus niger using phylogeny and comparative transcriptomics.* **Frontiers in Microbiology**, 2018. 9: p. 1045.
90. Hayer, K., M. Stratford, and D.B. Archer, *Structural features of sugars that trigger or support conidial germination in the filamentous fungus Aspergillus niger.* **Applied and Environmental Microbiology**, 2013. 79(22): p. 6924-6931.
91. Hayer, K., M. Stratford, and D.B. Archer, *Germination of Aspergillus niger conidia is triggered by nitrogen compounds related to L-amino acids.* **Applied and Environmental Microbiology**, 2014. 80(19): p. 6046-6053.
92. Fekete, E., et al., *D-Galactose uptake is nonfunctional in the conidiospores of Aspergillus niger.* **FEMS Microbiology Letters**, 2012. 329(2): p. 198-203.
93. Moore, B.M., et al., *Robust predictions of specialized metabolism genes through machine learning.* **Proceedings of the National Academy of Sciences**, 2019. 116(6): p. 2344-2353.
94. Yamanishi, Y., J.-P. Vert, and M. Kanehisa, *Supervised enzyme network inference from the integration of genomic data and chemical information.* **Bioinformatics**, 2005. 21(suppl_1): p. i468-i477.
95. Peng, M. and R.P. de Vries, *Machine learning prediction of novel pectinolytic*

*enzymes in Aspergillus niger through integrating heterogeneous (post-) genomics data.* **Microbial Genomics**, 2021. 7(12): p. 000674.

96. Todd, R.B., et al., *Prevalence of transcription factors in ascomycete and basidiomycete fungi.* **BMC Genomics**, 2014. 15(1): p. 1-12.

97. Benocci, T., et al., *Regulators of plant biomass degradation in ascomycetous fungi.* **Biotechnology for Biofuels**, 2017. 10(1): p. 1-25.

## Supplementary materials

**Figure S1.** Heatmaps visualization of expression profiles of sugar metabolism-related genes in six fungi during their growth on diverse monosaccharides. ***Available upon request from the author***

**Figure S2.** Heatmap of Pearson correlation of protein abundances of all proteins and sugar metabolism-related proteins in five fungi. ***Available upon request from the author***

**Figure S3.** Protein abundances profile of the sugar metabolism-related proteins involved in each sugar metabolic pathway in *A. niger* (A), *P. subrubescens* (B) and *T. reesei* (C). ***Available upon request from the author***

**Figure S4.** Heatmap of metabolites abundance levels under different monosaccharide conditions in five fungi. ***Available upon request from the author***

**Figure S5.** Correlation between abundance of metabolites and sugar metabolism-related genes in *A. niger* (A), *P. subrubescens* (B) and *T. reesei* (C). ***Available upon request from the author***

**Table S1.** List of species used in this study. ***Available upon request from the author***

**Table S2.** Orthologs of sugar metabolic genes detected among six studied species. ***Available upon request from the author***

**Table S3.** The predicted sugar metabolic pathways and corresponding genes in studied species based on ortholog mapping and analysis of Pathway Tools. ***Available upon request from the author***

**Table S4.** Expression profiling of sugar metabolism-related genes in six fungi (used for Figure 3 and Figure S1). ***Available upon request from the author***

**Table S5.** Protein abundances profile (log2 scaled) of the sugar metabolism-related proteins in five fungi (used for Figure 4 and Figure S3). ***Available upon request from the author***

**Table S6.** Heatmap of metabolites abundance levels under different monosaccharide conditions in five fungi (used for Figure S4). ***Available upon request from the author***

**Table S7.** Correlation between the abundance of metabolites and sugar metabolism-related genes in five fungi (used for Figure 5 and Figure S5). ***Available upon request from the author***

## Abbreviations

ACO, aconitate hydratase; ATP, adenosine triphosphate; CIT, citrate synthase; DakA, dihydroxyacetone kinase; DAK, dihydroxyacetone kinase; DGD, D-galactonate dehydratase; ENO, enolase; FBP, D-fructose 1,6-phosphatase; GaaA, D-galacturonic acid reductase; GaaB, L-galactonate dehydratase; GaaC, 2-keto-3-deoxy-L-galactonate aldolase; GaaD, L-glyceraldehyde reductase; GalE, UDP-galactose 4-epimerase; GalM, mutarotase; GalK, galactokinase; GalT, D-galactose-1-phosphate uridylyl transferase; GEO, Gene Expression Omnibus; GFD, glycerol 3-phosphate dehydrogenases; GlcA, glycerol kinase; GldB, glycerol dehydrogenase; GLK, glucokinase; GND, 6-phosphogluconate dehydrogenase; GO, Gene Ontology; GOX, glucose oxidase; GPD, glyceraldehyde 3-phosphate dehydrogenase; GPP, glycerol 1-phosphatase; GSD, glucose 6-phosphate dehydrogenase; GukA, gluconate kinase; GutB, D-sorbitol dehydrogenase; HXK, hexokinase; KEGG, The Kyoto Encyclopedia of Genes and Genomes; LadA, L-arabitol-4-dehydroge; LarA, NADPH-dependent L-arabinose reductase; LadB, galactitol dehydrogenase; LkaA, 2-keto-3-deoxy-L-rhamnonate aldolase; LraA, NADH-dependent L-rhamnose-1-dehydrogenase; LrdA, L-rhamnonate dehydratase; LrlA, L-rhamnonic acid lactonase; LXR, L-xylulose reductases; MGT, mannose-1-phosphate guanylyltransferase; NADH, nicotinamide adenine dinucleotide; OAH, oxaloacetate acetylhydrolase; PCP, pentose catabolic pathway; PFK, phosphofructokinase; PGDB, Pathway Genome Database; PGM, phosphoglycerate mutase; PGI, glucose 6-phosphate isomerase; PGK, phosphoglycerate kinase; PGL, 6-phosphogluconolactonase; PgmB, phosphoglucomutase; PKI, pyruvate kinase; PMI, mannose 6-phosphate isomerase; PMM, PPP, pentose phosphate pathway; PYC, pyruvate carboxylase; RBK, ribokinase; RBT, ribulokinase; RPE, ribulose-phosphate 3-epimerase; RPI, ribose 5-phosphate isomerase; SdhA, sorbitol dihydrogenase; SMGs, sugar metabolism-related genes; TAL, transaldolase; TCA, tricarboxylic acid; TKT, transketolases; TPI, triphosphate isomerase; XdhA, xylitol dehydrogenase; XhrA, L-xylo-3-hexulose reductase; XkiA, D-xylulose kinase; XyrA/XyrB, D-xylose reductase.

# Chapter 4

# Comparative genomics and transcriptomics analyses reveal divergent plant biomass-degrading strategies in fungi

This chapter was published in *Journal of Fungi*

Jiajia Li, Ad Wiebenga, Anna Lipzen, Vivian Ng, Sravanthi Tejomurthula, Yu Zhang, Igor V. Grigoriev, Mao Peng and Ronald P. de Vries

Volume 9, Pages 860, August 2023

DOI: https://doi.org/10.3390/jof9080860

## Abstract

Plant biomass is one of the most abundant renewable carbon sources, which holds great potential for replacing current fossil-based production of fuels and chemicals. In nature, fungi can efficiently degrade plant polysaccharides by secreting a broad range of carbohydrate-active enzymes (CAZymes), such as cellulases, hemicellulases and pectinases. Due to the crucial role of plant biomass degrading (PBD) CAZymes in fungal growth and related biotechnologies applications, investigation of their genomic diversity and transcriptional dynamics has attracted increasing attention.

In this project, we systematically compared the genome content of PBD CAZymes in six taxonomically distant species, *Aspergillus niger, Aspergillus nidulans*, *Penicillium subrubescens*, *Trichoderma reesei*, *Phanerochaete chrysosporium*, and *Dichomitus squalens*, as well as their transcriptome profiles during fungi grown on nine monosaccharides. Considerable genomic variation and remarkable transcriptomic diversity of CAZymes was identified, implying the preferred carbon source of these fungi and different transcription regulation in them. In addition, the specific carbon utilization ability inferred from genomics and transcriptomics was compared with the fungal growth profile on corresponding sugars, which to improve our understanding of the conversion process. This study enhances our understanding of genomic and transcriptomic diversity of fungal plant polysaccharide degrading enzymes, and provides new insights into designing enzyme mixtures and metabolic engineering of fungi for related industrial applications.

## 1. Introduction

Plant biomass is a magnificent renewable carbon source on Earth, which is of major importance for ecology and is attractive substrate for the production of biofuels and biochemicals. Plant cell walls mainly consist of polysaccharides (e.g., cellulose, hemicellulose and pectin), lignin and small amounts of protein. In addition, starch and inulin are commonly found as storage polysaccharides.

In nature, fungi play a central role in the degradation of plant biomass due to their unique ability to secrete a broad range of extracellular enzymes for decomposition of recalcitrant plant polysaccharides into mono- or oligosaccharides that can be metabolized for their growth and reproduction. Based on sequence and structural similarity, carbohydrate active enzymes are catalogued into five major families in the Carbohydrate-Active enZymes (CAZymes) database (http://www.cazy.org/) [1, 2], including glycoside hydrolases (GHs), carbohydrate esterases (CEs), polysaccharide lyases (PLs), glycosyltransferases (GTs) and auxiliary activities (AA). Fungal plant biomass degradation (PBD) has been extensively studied for several decades due to its essential roles in global carbon cycles and increasing potential in advancing the transition of fossil-based economy into more sustainable bio-based economy. The plant biomass degrading enzymes from these fungi have been widely used in many industrial sectors, such as biofuel, biochemical, paper and pulp, food, feed, detergents, and textiles [3].



Since polysaccharide composition is highly variable between different plant sources and environmental conditions, fungi need a highly diverse set of enzymes to degrade them (Table S1). To control their production, fungi have evolved a sophisticated regulatory system to tailor the energy-costly production and export of CAZymes for efficiently utilizing each specific plant substrate [4]. The outcome of this evolutionary adaptation is that different fungal species showed enormous diversity in terms of their CAZyme content, enzymatic production and transcriptional regulation related to plant biomass degradation. The importance of the regulatory layer is highlighted for example

by comparative genomics at the genus or section levels of *Aspergillus* revealed that fungal growth profiles on polysaccharides do not necessarily correlate with their PBD abilities inferred from genomic CAZyme content [5, 6]. An integrative analysis of genomic content, enzyme activity and proteomics data of eight taxonomically close *Aspergillus* species suggested that fungi with similar genomic potential for PBD could differ dramatically in CAZymes production and overall enzyme activities [7]. Similar results have been reported for *Trichoderma* and *Penicillium* species [8-11]. In addition, different species employ different sets of transcription factors (TFs) to regulate the utilization of specific polysaccharides. For instance, L-arabinose release from plant biomass and subsequent intracellular metabolism was controlled by AraR in the Eurotiomycetes [12], but the regulation was governed by a different unrelated TF, ARA1 in the Sordariomycetes and Leotiomycetes [13, 14]. More strikingly, although a similar set of the CAZymes were involved in PBD of Basidiomycete and Ascomycete fungi, very few of the characterized ascomycete regulators have corresponding one-to-one orthologous genes in basidiomycetes [4, 15].

Fungal enzymatic saccharification of plant polysaccharides results in the release of monomeric and oligomeric sugars, and in turn these small sugars or their metabolic conversion products act as inducer or repressor molecule for boosting or suppressing the production of PBD-related CAZymes [16-19]. For instance, D-galacturonic acid and L-rhamnose, the main monomeric sugar constituents of pectin, have been shown to induce the expression of pectinolytic enzymes in *Aspergillus niger*, *Trichoderma reesei*, and *Neurospora crassa* [20-23]. In *A. niger*, the expression levels of xylanases are controlled by two major regulators: induction via transcriptional activator XlnR and repression via transcriptional repressor CreA, depending on different xylose concentration [17]. However, most of previous (post-)genomics studies on fungal PBD-related CAZymes were restricted to model species and were performed on limited growth conditions, or solely relied on comparative genomics analysis.

New studies based on both genomics and post-genomics on taxonomically distant species will enhance our understanding of the molecular mechanisms and biodiversity of fungal PBD, which will provide new insights into the design of enzyme mixtures and guide metabolic engineering of fungi towards improved plant biomass conversion during the related industrial applications.

In this study, we systematically compared the genomic repertoires of PBD-related CAZymes in six evolutionarily diverse species, *A. niger*, *Aspergillus nidulans*, *Penicillium subrubescens*, *T. reesei*, *Phanerochaete chrysosporium*, and *Dichomitus squalens*, as well as their transcriptomic response to nine plant-derived monosaccharides. In addition, the specific carbon utilization ability inferred from genomics and transcriptome data was compared to fungal growth profiles across a wide range of carbon sources.

## 2. Materials and Methods



### *2.1. Fungal strains*

This study focuses on six evolutionarily diverse species: three Eurotiomycetes (*A. niger* N402, *A. nidulans* FGSCA4 and *P. subrubescens* CBS132785/FBCC1632), one Sordariomycete (*T. reesei* QM6a) and two Basidiomycetes (*P. chrysosporium* RP-78 and *D. squalens* FBCC312/CBS432.34). The genomes and related annotation of these six fungal species were described in our previous study [24].

### *2.2. Identification of orthologs of CAZyme genes related to plant biomass degradation across six fungal species*

CAZyme annotation and related protein sequences of each species was obtained from JGI MycoCosm Portal (https://mycocosm.jgi.doe.gov). The PBD-related CAZy families (based on previous studies) were selected for further analysis [7, 25, 26]. Orthologous genes were identified using OrthoMCL [27] (https://orthomcl.org/orthomcl/app) and OrthoFinder [28] (https://github.com/davidemms/OrthoFinder) with the protein sequences as input and default parameters applied. We took the combination of the two

ortholog mapping results as our final ortholog set.

### *2.3. Phylogenetic analyses*

Protein sequences of each selected CAZy family (PBD related) were aligned by MAFFT v7.310 [29, 30]. Unusually long and incomplete sequences were corrected manually based on NCBI BlastX searching results, whereas duplicate and ambiguous sequences were discarded. Phylogenetic trees generated using FastTree [23] were used to distinguish true orthologs across six fungi. The resulting trees were visualized using iTOL [31].

### *2.4. Transcriptome sequencing and analysis*

The transcriptome data of six monosaccharides of *D. squalens* was obtained from our previous study (Gene Expression Omnibus (GEO) database accession: GSE105076) [18].

Transcriptome data of *A. niger* was newly generated in this study and was deposited in the Sequence Read Archive at NCBI under the accession numbers: SRP448993, SRP449003-SRP449007, SRP449023, SRP449039, SRP449049,SRP449062,SRP449079-SRP449081,SRP449083-SRP449085, SRP449089, SRP449068-SRP449070, SRP449098, SRP449125, SRP449138, SRP449141, SRP449142, SRP449151, SRP449193, while the other four species grown on nine monosaccharides were recently generated in our previous study [24].

The culture conditions were maintained as similar as possible to allow a good comparison of the species. In detail, *A. niger* NRRL3, *A. nidulans* FGSC A4 and *P. subrubescens* FBCC1632/CBS132785 were pre-cultured in complete medium [32] with 2% D-fructose and mycelial aliquots were then transferred to minimal medium [32] with 25 mM D-glucose, D-fructose, D-galactose, D-mannose, L-rhamnose, D-xylose, L-arabinose, D-galacturonic acid and D-glucuronic acid, respectively, and cultivated for 2 h. The same cultivation approach was used for *T. reesei* QM6a and *P. chrysosporium* PR-78, but with media optimized for these species [33, 34], and cultivated for 4 h. The

cultivation approach for *D. squalens* can be found in our previous study [18]. So only the basic salt medium is specific for the species, but the carbon source is the same in all species. Mycelial samples were harvested and immediately frozen in liquid nitrogen.

Total RNA was extracted from ground mycelial samples using TRIzol reagent (Invitrogen). Purification of mRNA, synthesis of cDNA library and sequencing were conducted at DOE Joint Genome Institute (JGI). The details of transcriptome data generation and related downstream analysis were described previously [24].

In addition, the CAZy genes with strong sugar specificity were selected by a sugar specificity index (*SSI*), which was calculated by following the equation that has been widely used in studies of gene expression specificity [35, 36]:

$$SSI = \frac{\sum_{i=1}^{n}(1-\hat{x_i})}{n-1}; \hat{x_i} = \frac{x_i}{\max\limits_{1\le i\le n}(x_i)}$$



For the above equation, $x_i$ is the expression of the gene in each specific monosaccharide condition *i*, and n is the number of total tested monosaccharide conditions. We define the CAZyme genes with *SSI* > 0.7 as sugar specific genes (SSGs). In addition, we filtered out the extremely low expressed genes whose maximum expression levels (FPKM) were less than 10 in all tested conditions. For each SSG, we defined the corresponding sugar growth condition in which the gene showed highest expression level among all tested sugars as its inducing sugar.

### *2.5. Growth profiling on 18 plant biomass-related substrates*

For growth phenotype analyses, all strains were grown on minimal medium (MM) [32] with monosaccharides, and disaccharides at 25 mM and polysaccharides at 1% final concentration. Growth was performed at 30°C for two *Aspergillus* spp. and 25°C for the other species. Media with no carbon source was used as a control. If growth on a specific carbon source is the same as on no carbon source, it is considered as no growth. Growth

on D-glucose was used as an internal reference for growth. Substrates were obtained from Sigma-Aldrich (Zwijndrecht, The Netherlands): D-glucose (G-8270), D-fructose (F-0127), D-galactose (G-0625), D-mannose (M-2069), L-rhamnose (83650), D-galacturonic acid (48280), D-glucuronic acid (G-8645), sucrose (S-5016), maltose (M-9171), beechwood xylan (X-4252), guar gum (G-4129), apple pectin (76282), inulin (I-2255), cellulose (C8002); Difco (Leeuwarden, The Netherlands): soluble starch (217820); or Acros (Geel, Belgium): cel-lobiose (108465000), citrus pectin (416862500).

## 3. Results

### *3.1. Genomic potential of PBD CAZymes shows diversity across six fungi*

In total, the genome repertoire of 65 different CAZy (sub-)families and feruloyl esterases (FAEs) involved in degradation of different plant polysaccharides was comparatively analyzed in six fungi (Figure 1 and Figure S1). Specifically, the selected PBD-related enzymes studied included four auxiliary activities (AA) families (AA3, AA9, AA13, AA16), six carbohydrate esterase families (CE1, CE5, CE8, CE12, CE15, CE16), five polysaccharide lyases families (PL1, PL3, PL4, PL9, PL11), 39 glycoside hydrolases families (GHs) and FAEs (Figure 1). Notably, ten enzyme families (GH3, GH12, CE16, GH2, GH35, GH31, GH43, GH51, GH54, FAE) are multifunctional and involved in the degradation of more than one plant polysaccharide (Figure S1).

The potential for PBD as inferred from the genomic repertoire PBD CAZymes shows clear diversity across the studied fungi (Figure 1 and Table S2). Overall, the CAZyme repertoire differences correlate well with the taxonomic distance of the studied fungi. The three Eurotiomycetes (*A. niger, A. nidulans and P. subrubescens*) possess a relative high number of PBD related CAZymes compared to *T. reesei* and the Basidiomycetes (*P. chrysosporium* and *D. squalens*), especially with respect to enzymes from GH and PL families (Figure S2 and Table S3). *T. reesei* contains less CAZymes involved in each plant polysaccharide then the other three Ascomycetes.

The two Basidiomycetes have less hemicellulolytic, pectinolytic and starch degrading enzymes, while the total numbers of cellulolytic enzymes are comparable to Eurotiomycetes. Notably, a higher number of AA9 genes were identified in the Basidiomycetes. In the following sections, we discussed in more detail the diversity of genomic potential with respect to degradation of each polysaccharide.

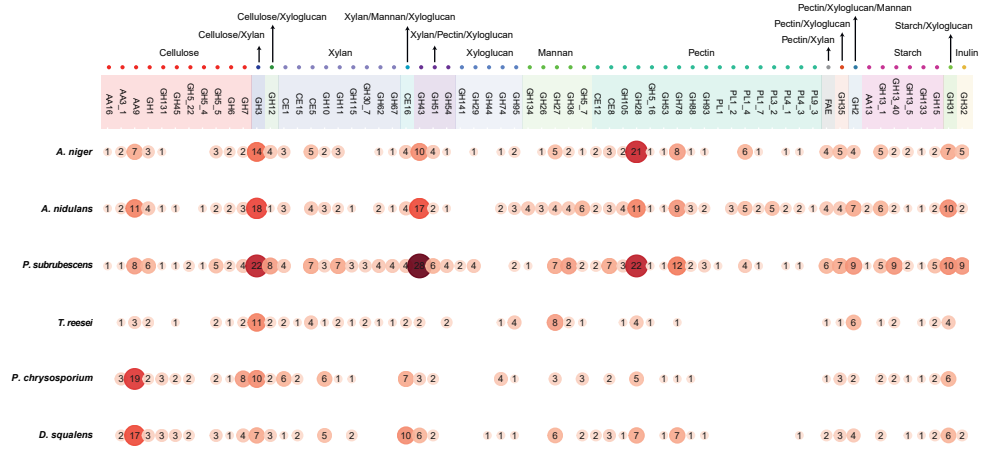


**Figure 1**. **Distribution of gene numbers in selected CAZy families in the six fungi**. Bubbles with numbers represent the number of encoding genes in each CAZy family. Color intensity and bubble size are related to the numbers within them.

### *3.1.1. Cellulose degradation*

Cellulose is a major component of plant biomass. The degradation of cellulose mainly involves in four groups of enzymes: endoglucanases (EGLs), cellobiohydrolases (CBHs), and β-glucosidases (BGLs), as well as auxiliary activities enzymes (AAs) (Table S1) [37].

EGLs, act mostly on amorphous regions of cellulose, releasing reducing and nonreducing chain ends [38, 39]. They are mainly classified in GH12, GH131, GH45 and three subfamilies of GH5 (GH5_4, GH5_5 and GH5_22) (Table S3). Many of these EGLs families showed noticeable difference between both evolutionarily close and distant species, especially the GH5 (sub-)family. GH5_4 has a single copy in *A. nidulans* and *P. subrubescens*,

but missing in all other species. GH5_22 was not identified in two *Aspergillus* and *T. reesei*, but present as two copies in *P. subrubescens* and the two Basidiomycetes. *P. subrubescens* has more copies of GH5_5 and GH12 than other species. In contrast to most EGLs with comparable or more copies in Ascomycetes than Basidiomycetes, GH45 and GH131 have multiple copies in both Basidiomycetes, but were present as a single copy in three Ascomycetes (except for the absence of GH131 in *T. reesei*).

CBHs release cellobiose from the reducing (CBH I) or non-reducing (CBH II) ends of the cellulose. Fungal CBH I and CBH II genes are classified in GH7 and GH6, respectively. Gene copies of GH7 are notably larger (eight) in *P. chrysosporium* and varying among the other species (two to four). GH6 was present as a single copy in *T. reesei* and the Basidiomycetes, and with two genes in the three Eurotiomycetes.

BGLs are involved in the cleavage of cellobiose into glucose, and are mainly grouped into the GH1 and GH3 families. GH3 BGLs are greatly expanded in the three Eurotiomycetes, while GH1 is only significantly expanded in *P. subrubescens*.

LPMOs and cellobiose dehydrogenases (CDHs) have been suggested to be involved in the oxidative cleavage of cellulose and boosted cellulose degradation [26, 40, 41]. AA9 is a widely distributed LPMO in fungi, which was present as multiple copies in all studied fungi, and was significantly expanded in the Basidiomycetes compared to the Ascomycetes. In contrast, AA16 LPMO was only present as a single copy in the Eurotiomycetes and was missing in the other species, while AA3_1 CDH showed less variation among the studied fungi.

*3.1.2. Hemicellulose degradation*

In contrast to cellulose, hemicellulose is a heterogeneous polysaccharide, which is composed of different residues that form diverse backbones and branches. Degradation of plant hemicelluloses requires a variety of

hemicellulolytic enzymes, which are divided over at least 20 GH families, four CE families and FAE (Figure1 and Figure S1 and Table S1).

Xylan degradation

Xylan is a common hemicellulose polysaccharide. To degrade xylans, a specific set of CAZymes is required, mainly including: β-D-endoxylanase (XLN) from GH10 and GH11 [42], β-xylobiohydrolase (XBH) from GH30_7, BXL from GH3 and GH43, α-L-arabinofuranosidase (ABF) from GH43, GH51 and GH54, α-glucuronidase (AGU) from GH67 and GH115, glucuronoyl esterase (GE) in CE15, acetyl xylan esterase (AXE) from CE1 and CE5, hemicellulose acetyl esterase (HAE) from CE16 [43], arabinoxylan arabinofuranohydrolases (AXH) from GH62, and feruloyl esterase (FAE). Overall, xylanolytic enzyme encoding genes in the six studied species differs significantly. The genome of *P. subrubescens* contains the largest number of xylanases, followed by the two *Aspergillus* species, while *T. reesei* and the two Basidiomycetes had the lowest number. Higher gene copies in GH3, GH43, GH51 and GH54 families were found in the Ascomycetes than the Basidiomycetes, while GH10 and CE16 showed opposite trends. In addition, CAZy families GH54 (ABF), GH62 (AXH), GH67 (AGU) and CE5 (AXE) were exclusively found in the Ascomycetes. CE15 is absent in the Eurotiomycetes but present in the two Basidiomycetes and *T. reesei*, while more copies of FAEs and GH43 were identified in Eurotiomycetes than in the other species.

Xyloglucan degradation

Xyloglucan has complex structure with various sidechain and decorations [44], thus its hydrolysis requires a vast arsenal of enzymes. Except for ABF and HAE enzymes mentioned earlier, xyloglucan degrading enzymes also include α-fucosidase (AFC) from GH29, GH95 [45] and GH141, xyloglucan β-1,4-endoglucanase (XEG) from GH12 [46], GH44 [47] and GH74 [48], α-xylosidase (AXL) from GH31 [45], and β-1,4-galactosidase (LAC) from GH2 and GH35 [37]. High variability was observed in the copy number

of these CAZy families. *P. subrubescens* has expanded several xyloglucan degrading enzyme families, such as GH12, GH43 and GH54. Numbers in families GH95, GH12, GH35, GH2 and GH31 varied remarkably in the Ascomycota, but showed low variation between the two Basidiomycetes (Figure 1). In contrast, GH29 was exclusively identified in *A. niger* and *P. subrubescens,* while GH44 and GH141 were only identified in *D. squalens* and *P. subrubescens*, respectively.

Mannan degradation

The degradation of mannans requires the concerted action of several mannanases, including β-1,4-endomannanase (MAN), β-1,4-mannosidase (MND), α-galactosidases (AGL) and HAE. Fungal AGLs have been assigned to GH27 and GH36. MANs are mainly classified in GH5_7 with fewer copies in GH26 and GH134. MND and HAE belong to GH2 [49] and CE16, respectively (Figure 1).

Overall, the genomes of *A. nidulans* and *P. subrubescens* contain more copies of mannanases than the other fungi (Figure 1). The differences are that *A. nidulans* has more copies of MAN (GH26, GH134 and GH5_7), whereas *P. subrubescens* harbors large number of AGL (GH27 and GH36) and MND (GH2) encoding genes. In general, the other four species have less mannanases, except that *T. reesei* and *A. niger* contain considerable number of AGL (especially GH27) and MND, and the Basidiomycetes have a higher number of CE16 copies. Notably, GH36, GH26 and GH134 were absent in both Basidiomycetes (Figure 1), and GH26 and GH134 were absent in the Basidiomycetes and *T. reesei*.

*3.1.3. Pectin degradation*

Pectin is a very complex polymer, composed of a family of galacturonic acid-rich polysaccharides, including: homogalacturonan (HG), xylogalacturonan (XGA), and two types of rhamnogalacturonan (RG-I and RG-II) [50]. Its degradation requires a variety of enzymes (Table S1), which are classified

into more than 20 CAZy families, mainly including pectin hydrolases (GH2, GH5_16, GH28, GH35, GH43, GH51, GH54, GH78, GH88, GH93 and GH105), pectin lyases (PL1, PL3, PL4, PL9 and PL11) [51], and pectin methyl esterases (CE8), and pectin and rhamnogalacturonan acetyl esterases (CE12) [52]. The gene content of pectinolytic genes showed huge difference among six studied species. In general, three Eurotiomycota fungi have higher number and more diverse set of pectinolytic genes than other three studied species (Figure 1). Several enzyme families involved in pectin degradation are reduced or absent in *T. reesei*, including the absence of four GH families (GH51, GH53, GH88, and GH93) and the much less genes encoding for GH28 and GH78 compared to other three Ascomycetes. PLs are exclusively present and largely varied in three Eurotiomycota species, of which *A. nidulans* has more types of PL families (e.g. PL1, PL3, PL4, PL9 and PL11). The other studied species loss most of the PLs, except that *D. squalens* has one gene of PL4_3. Regarding pectin methylesterases, CE8 was found as multiple copies in most species except the absence in *T. reesei,* while CE12 was present in the three Eurotiomycota species and *D. squalens* and absence in *T. reesei* and *P. chrysosporium*.



*3.1.4. Storage polysaccharides degradation*

Starch and inulin are two main storage polysaccharides of plant. Starch consists of an α-1,4-linked polymer (amylose) of D-glucose residues which can be branched at α-1,6-linked points (amylopectin) [53-55]. It is degraded by the joint enzymatic actions of α-glucosidase (AGD) from GH13_40 and GH31, amylo-α-1,6-glucosidase (AMG) from GH133, α-amylase (AMY) from GH13_1 and GH13_5, glucoamylase (GLA) from GH15 and LPMOs from AA13 (Table S1). Most starch degrading CAZymes were identified with one or more copies in all studied species, except that AA13 was only identified in *A. nidulans* and *P. subrubescens,* and the absence of GH13_5 and GH13_40 in *T. reesei* and *D. squalens*, respectively. Additionally, the three Eurotiomycetes contain more copies of GH13_1 than the other species, while

*P. subrubescens* harbors the largest number of GH13_40 and GH15 (Figure 1).

Inulin consists of a branched β-2,1-linked chain of D-fructose with a terminal D-glucose residue [56, 57]. GH32 is the only CAZy family involved in inulin degradation, which mainly acts as endo-inulinase, exo-inulinase and invertase. GH32 enzymes were absent in *T. reesei* and *P. chrysosporium*, but were expanded in *A. niger* and *P. subrubescens* with five and nine copies, respectively (Figure 1).

### *3.2. Expression profile of CAZy genes during fungi grown on different monosaccharides*

PBD CAZymes mediate the release of monomeric/oligo- sugars from plant polysaccharides. In turn, the presence of these small sugars can induce or repress the expression of CAZymes encoding genes. Here, we investigated the transcriptome diversity of CAZymes in response to different plant-derived monosaccharides of studied fungi.

Overall, the expression profiles of CAZy genes varied across the tested species and enzyme families. The inducing pattern of pectinolytic genes showed considerable consistency across the studied species (Figure 2). For most of the studied Ascomycetes, relative high abundance of pectinolytic genes were observed on L-arabinose, D-galacturonic acid, D-glucuronic acid or L-rhamnose, which are the most abundant sugar present in the polymers of pectin. *A. nidulans* is an exception, in which pectinolytic genes were mainly higher expressed on L-rhamnose, and were not clearly induced by D-galacturonic acid (Figure 2 and Table S4). Additionally, the genes involved in the degradation of xylan and xyloglucan showed relatively high expression on L-arabinose in the Ascomycetes, particularly *P. subrubescens*, which had more highly expressed genes on L-arabinose, while they were also induced by D-xylose in *A. niger* and induced by L-rhamnose in *A. nidulans*. Notably, no clear sugar inducing pattern was observed for the studied Basidiomycetes, except that most of PBD-related CAZymes showed relatively high expression

levels on D-galacturonic acid compared to the other tested sugars for *P. chrysosporium*.

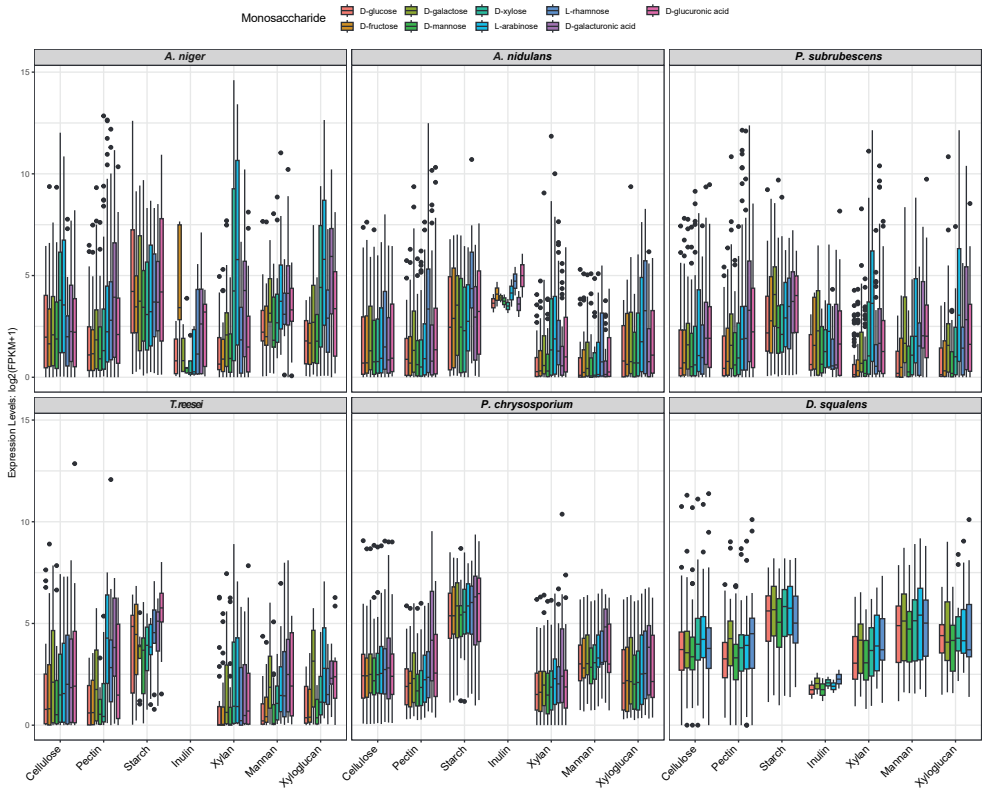


**Figure 2. Distribution of the expression profiling of the CAZy genes involved in degradation of different polysaccharides (mentioned at the bottom) in the tested fungi**. On the top, different colors indicate different monosaccharides that were used as carbon source in the fungal cultures. The black dots indicate outlier genes with extremely low or high expression values.

In addition, we also analyzed sugar specificity index of each CAZy gene. Sugar specificity is often described based on gene expression levels, and we define that the genes with high sugar specificity index are a group of genes whose function and expression are specific for one certain sugar condition. This index can be used to evaluate if the specific CAZy genes involved in certain polysaccharide degradation showed corresponding inducing patterns on the corresponding monosaccharides (Figure S3 and Table S5). As

expected, in the three Eurotiomycetes, the CAZy genes involved in pectin degradation showed clear D-galacturonic acid and L-rhamnose inducing specificity (except mainly L-rhamnose inducing patterns in *A. nidulans*), and the CAZy genes involved in xylan or xyloglucan degradation revealed L-arabinose and D-xylose inducing patterns. Notably, parts of CAZy genes involved in the degradation of the above polysaccharides also showed high expression levels on D-galactose in *A. niger* and *A. nidulans*, but not in *P. subrubescens*, which indicates metabolic cross talk between sugar-sensing pathways and/or overlap in regulatory networks. In addition, ten and 18 CAZymes involved in pectin, xylan or mannan degradations showed sugar specific induction on D-glucuronic acid in *A. nidulans* and *P. subrubescens*, respectively. In contrast, *T. reesei* only had a small number of CAZy genes with a high sugar specificity index, such as six CAZy genes involved in xylan and xyloglucan that showed L-arabinose induction, and three pectinolytic genes with D-galacturonic acid induction specificity (Table S5). We detected more than twenty CAZymes highly expressed on D-galacturonic acid in *P. chrysosporium*, but only nine genes belong to pectinolytic genes. In *D. squalens*, due to the limited transcriptome data, we only observed very small number of genes showing sugar specificity (Figure S3 and Table S5).

### *3.3. Growth profiles are not always correlated with (post-)genomics profile of CAZy*

The noticeable diversity of genomics and expressional profiles of CAZy genes identified above in the tested fungi strongly suggests that they employ different PBD strategies. To evaluate the correlation between (post-)genomic profiles and actual carbon utilization ability, we analyzed the growth profile of these fungi on nine monosaccharides, two disaccharides and seven polysaccharides (Figure 3A). Growth on D-glucose and no carbon source were used as an internal reference for growth. The relative difference between growth on different carbon sources was examined between the species.

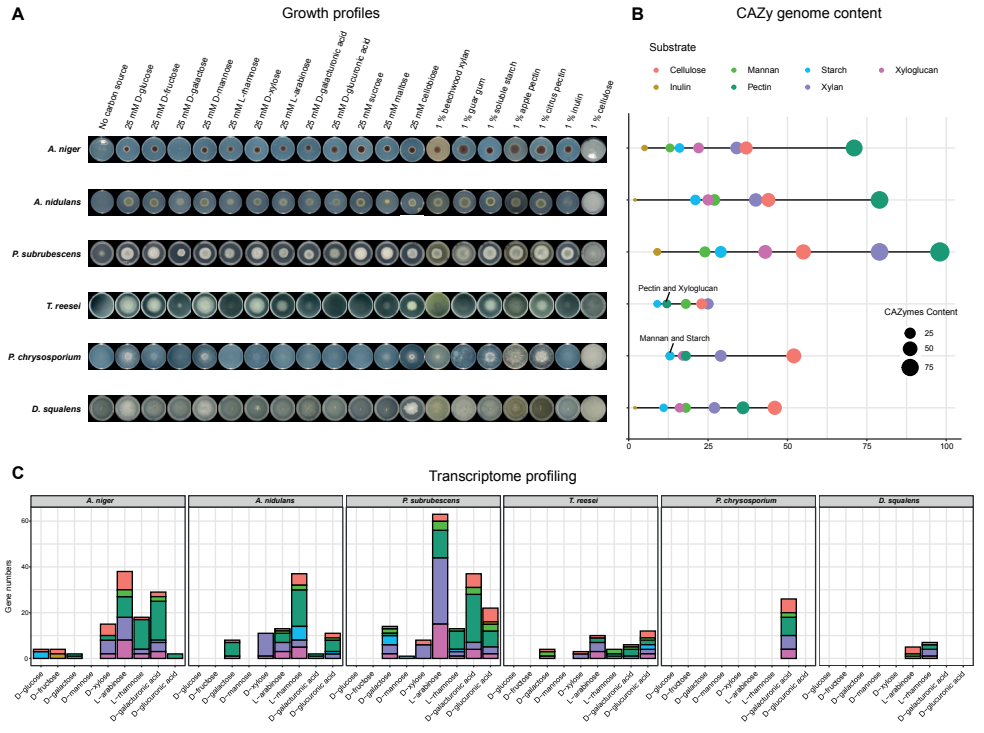


**Figure 3. Overview of growth profiles, CAZyme content and transcriptome profiling in the studied fungi**. **A**. Growth profiles on various plant biomass related carbon sources. **B**. Summary of CAZy genome content involved in the degradation of each plant polysaccharide. The size of the dots indicates the number of CAZy genes, and the color indicates different substrates that CAZymes act on. **C**. Distribution of CAZy genes with sugar specific expression on each monosaccharide. The same color code as Figure 3B was used to indicate the different substrates.

In general, the Eurotiomycetes showed better growth on most of the tested polysaccharides than the other species, which is consistent with a higher gene content of the corresponding degrading genes (Figure 3B). In contrast, possibly due to the absence and reduction of many CAZy genes in the genome, *T. reesei* grew relative poorly on most of the tested polysaccharides (except starch). Different species showed huge variation of growth on inulin. *A. niger* and *P. subrubescens* grew well on inulin, whereas the other four species grew very poorly. This correlates well with the diversity of inulin degrading genes in different species. *A. niger* and *P. subrubescens* have many

copies of inulinolytic genes, in contrast to the other species that are impaired in their inulin degrading machinery. Compared to the other two analyzed Eurotiomycetes, *A. nidulans* had less copies of GH32, particularly missing copies of endo-inulinase (Table S3), which reduces its ability to efficiently utilize this substrate as a carbon source. In *T. reesei* and *P. chrysosporium*, GH32 was fully absent. In *D. squalens*, two copies of GH32 were identified, but they were extremely lowly expressed (Table S4).

Besides these consistent observations for genomic repertoire, expression and growth, also inconsistencies were observed. Good growth on citrus pectin was observed for *T. reesei* and *P. chrysosporium*, while their genome encodes fewer pectinolytic genes than the Eurotiomycetes. Additionally, we observed active gene expression of pectinolytic genes for these two species on D-galacturonic acid, the major sugar component of pectin (Figure 2 and Figure 3C). In contrast, *D. squalens* grew poorly on pectin. In line with the crucial role of pectin lyases (PLs) on degrading apple pectin [58], the absence of PLs on the genome of *T. reesei* and *P. chrysosporium* could explain their relative poor growth on apple pectin. All tested fungi could grow relative well on starch, while none of the studied species grew well on cellulose as the sole carbon source despite a significant number of cellulases encoded in their genomes. The Eurotiomycetes grew well on xylan, while the other three tested species showed only minimal growth on xylan, which is consistent with the fact that fewer xylanase genes were identified in their genomes and actively expressed during growth of these fungi on xylan-related monosaccharides (Figure 3). In contrast, we observed poor correlation between (galacto-)mannan (guar gum) and mannan degrading CAZymes content in the fungal genomes. The Eurotiomycetes grew well on guar gum, even though *A. niger* has a very small number of mannan degrading genes. Compared to *A. niger*, *T. reesei* and the Basidiomycetes have a comparable number of mannan degrading genes, but they all showed impaired growth. Unexpectedly, none of the fungi showed expression specificity on mannose (Figure 3C and Table S5), but showed high

sugar specificity on other monosaccharides, like *A. niger* and *P. subrubescens* on L-arabinose. For *A. niger* it has been shown that mannobiose, rather than mannose is the inducer of the mannanolytic enzyme system [59], which could explain why there is no mannose-specific induction observed.

## 4. Discussion

The genome content of CAZymes showed considerable differences among the tested species, especially among the taxonomically distant species. The CAZymes repertoire and their expression under different carbon source of each specific fungus implies the possible different PBD approaches. In general, the studied Eurotiomycetes possess more diverse sets of PBD genes that enable them to grow well on most tested polysaccharides, while *T. reesei* and two Basidiomycetes grew poorly on most polysaccharides due to relatively small repository or less actively expressed PBD genes. Particularly, *T. reesei* is the only tested Ascomycete that showed impaired growth on maltose and starch, which correlates with less starch degrading enzymes identified in the genome of this species.



Among the three tested Eurotiomycota, only *A. niger* and *P. subrubescens* can grow well on both sucrose and inulin, while *A. nidulans* showed impaired growth on inulin. The divergence could be explained by the lower copies of GH32, and the low expression level of intracellular invertase SucB (Aspnid1|2301) [60].

The detailed structure and composition of polysaccharides significantly affects its degradation by specific fungi, especially for species with less variety of enzyme families encoded in their genome, which was demonstrated by the growth profile of *T. reesei* and *P. chrysosporium* on pectin extracted from two different plant sources.

Although crucial PBD CAZy genes were present in their genomes, the two tested wood-degrading Basidiomycetes showed relatively poor growth on most tested polysaccharides and showed clear different gene expression

profiles of PBD genes than the tested Ascomycetes. The possible reason could be that these fungi mainly grow on wood in nature, and have adapted their induction system to grow on complex polysaccharides as carbon sources instead of being optimized for individual polysaccharides as we tested here.

Poor growth on cellulose was observed for all tested species, although cellulolytic genes were identified in their genomes. One possible explanation is that the necessary enzymes were not fully activated during fungal growth on cellulose. For instance, many cellulolytic genes in *A. niger* are controlled by the xylanolytic transcriptional activator XlnR, which was mainly activated in the presence of xylose [61, 62]. In line with this, visible growths on the cellulose-derived oligosaccharide, cellobiose, were observed for all six species (Figure 3A), which require much fewer degrading enzymes to fully decompose than intact cellulose. In addition, commercial cellulose is not structurally identical to natural cellulose, due to the method of extraction.

We identified several CAZy families that are uniquely present in certain species or phyla, which could indicate a special physiological role and potential industrial applications. For example, the AA16 family only exists in the three Eurotiomycetes, while the AA9 family is broadly present in all six species. A recent study suggested these two fungal LPMO families were counterparts and targeted the same substrates [26]. Contradicting this, a recent study found that AA16 can only boost the cellulase activity of AA9, but cannot break down cellulose independently [63]. Clearly, more detailed studies in this area are warranted. In addition, among the six tested species, the GH44 enzymes were only identified in *D. squalens* [47], which indicates the scarcity and unique function of this enzyme. Additionally, we found CE15 only in the two Basidiomycetes and *T. reesei*, which was proposed to cleave ester linkages between lignin and glucuronoxylan [64]. GH141, whose active site pocket of the hydrolase is likely to be significantly more open than other fucosidases (e.g. GH95 and GH29), was only present in *P. subrubescens*, revealing it has a more variable PBD mode [65]. Only *A. nidulans* contains genes encoding

PL3 and PL9 pectate lyases, as well as PL11 rhamnogalacturonan lyases, which indicated that *A. nidulans* may have great potential for commercial applications in pectin degradation [66].

In addition, some genes encoding specific family present phyla specificity. GH54 is only present in the four Ascomycetes, and has been reported to remove decorations including L-arabinose and D-galactose in the presence of metal cofactors, which may suggest that it could be introduced in enzymatic cocktails to assist in the hydrolysis of xylan and xyloglucan. Similarly, GH62 was only found in the four Ascomycetes, and is the sole GH family that contains only arabinofuranosidases (ABFs) specificities [67]. However, high variability was also observed in the number of genes within the same phyla. For example, multiple gene copies of GH51, GH54 and GH62 encoding ABFs were identified in *P. subrubescens,* which implies a potential enhancement in the ability of *P. subrubescens* to release L-arabinose from plant biomass in a more specific manner [68]. Moreover, this is in line with our finding that CAZy genes involved in xylan, pectin and xyloglucan in *P. subrubescens* show evident L-arabinose specificity (Figure 3C).



Additionally, the expression profile of PBD genes in six tested species showed dramatic differences during fungal growth on nine monosaccharides (Figure 2 and Figure 3C). For the closely related Eurotiomycetes, although they have comparable content of pectinolytic genes, the varying expression patterns of these genes suggest the activity of an underlying complex regulatory networks. In *A. niger,* a comparable set of pectinolytic genes were induced on L-arabinose, L-rhamnose, and D-galacturonic acid, while the same set of genes in *A. nidulans* were largely induced by L-rhamnose alone, and in *P. subrubescens* mainly induced by L-arabinose and D-galacturonic acid (Table S5). Similarly, sugar-specific CAZymes involved in xylan or xyloglucan degradation in different fungi also displayed different inducing patterns on L-arabinose and D-xylose.

## 5. Conclusion

In this study, we revealed the diversity of fungal plant polysaccharide degrading enzymes at both genomics and transcriptomics level. Considering the importance ecological role and increasing industrial potential of fungal plant biomass degradation, the findings revealed in this study will improve our understanding of the fungal diversity and provides new insights into designing enzyme mixtures and metabolic engineering of fungi for related industrial applications.

## Author Contributions

All authors contributed to the manuscript. R.P.d.V. conceived and supervised the overall project. J.L. performed the formal analysis and wrote the original draft. A.V. performed the growth profiles. M.P. and R.P.d.V. revised the manuscript. S.T., Y.Z., A.L., V.N. and I.V.G. performed and coordinated the RNAseq experiments and initial analyses. All authors have read and agreed to the published version of the manuscript.

## Funding

J.L. was supported by the China Scholarship Council (CSC 201909110079).

## Data Availability Statement

The reads from each of the transcriptome sequencing (RNA-seq) samples were deposited in the Sequence Read Archive at NCBI under the accession numbers: *A. niger* SRP448993, SRP449003-SRP449007, SRP449023, SRP449039, SRP449049, SRP449062, SRP449079-SRP449081, SRP449083-SRP449085, SRP449089, SRP449068-SRP449070, SRP449098, SRP449125, SRP449138, SRP449141, SRP449142, SRP449151, SRP449193; *A. nidulans* SRP262827-SRP262853; *P. subrubescens* SRP246823-SRP246849; *T. reesei* SRP378720-SRP378745; *P. chrysosporium* SRP249214-SRP249240.

## Acknowledgments

This research was performed under the Facilities Integrating Collaborations for User Science (FICUS) program (proposal: 10.46936/fics.proj.2018.50379/60006403) and used resources at the DOE Joint Genome Institute (https://ror.org/04xm1d337) which is a DOE Office of Science User Facility operated under Contract Nos. DE-AC02-05CH11231.

## Conflicts of Interest

The authors declare no conflicts of interest.

## References

1. Cantarel, B.L., et al., *The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics.* **Nucleic Acids Research**, 2009. 37(suppl_1): p. D233-D238.
2. Lombard, V., et al., *The carbohydrate-active enzymes database (CAZy) in 2013.* **Nucleic Acids Research**, 2014. 42(D1): p. D490-D495.
3. Meyer, V., et al., *Growing a circular economy with fungal biotechnology: a white paper.* **Fungal Biology and Biotechnology**, 2020. 7(1): p. 1-23.
4. de Vries, R.P. and M.R. Mäkelä, *Genomic and postgenomic diversity of fungal plant biomass degradation approaches.* **Trends in Microbiology**, 2020. 28(6): p. 487-499.
5. Kjærbølling, I., et al., *A comparative genomics study of 23 Aspergillus species from section Flavi.* **Nature Communications**, 2020. 11(1): p. 1-12.
6. de Vries, R.P., et al., *Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus Aspergillus.* **Genome Biology**, 2017. 18(1): p. 1-45.
7. Benoit, I., et al., *Closely related fungi employ diverse enzymatic strategies to degrade plant biomass.* **Biotechnology for Biofuels**, 2015. 8(1): p. 1-14.
8. Rosolen, R.R., et al., *Whole-genome sequencing and comparative genomic analysis of potential biotechnological strains from Trichoderma harzianum, Trichoderma atroviride, and Trichoderma reesei.* **bioRxiv**, 2022: p. 2022.02. 11.479986.
9. Horta, M.A.C., et al., *Network of proteins, enzymes and genes linked to biomass degradation shared by Trichoderma species.* **Scientific Reports**, 2018. 8(1): p. 1341.
10. Lenz, A.R., et al., *Analysis of carbohydrate-active enzymes and sugar transporters in Penicillium echinulatum: A genome-wide comparative study of the fungal lignocellulolytic system.* **Gene**, 2022. 822: p. 146345.

11. Méndez-Líter, J.A., et al., *Hemicellulases from Penicillium and Talaromyces for lignocellulosic biomass valorization: A review.* **Bioresource Technology**, 2021. 324: p. 124623.
12. Battaglia, E., et al., *Analysis of regulation of pentose utilisation in Aspergillus niger reveals evolutionary adaptations in Eurotiales.* **Studies in Mycology**, 2011. 69(1): p. 31-38.
13. Klaubauf, S., et al., *A novel L-arabinose-responsive regulator discovered in the rice-blast fungus Pyricularia oryzae (Magnaporthe oryzae).* **FEBS Letters**, 2016. 590(4): p. 550-558.
14. Benocci, T., et al., *ARA 1 regulates not only L-arabinose but also D-galactose catabolism in Trichoderma reesei.* **FEBS Letters**, 2018. 592(1): p. 60-70.
15. Benocci, T., et al., *Regulators of plant biomass degradation in ascomycetous fungi.* **Biotechnology for Biofuels**, 2017. 10(1): p. 1-25.
16. Amore, A., S. Giacobbe, and V. Faraco, *Regulation of cellulase and hemicellulase gene expression in fungi.* **Current Genomics**, 2013. 14(4): p. 230-249.
17. de Vries, R.P., J. Visser, and L.H. de Graaff, *CreA modulates the XlnR-induced expression on xylose of Aspergillus niger genes involved in xylan degradation.* **Research in Microbiology**, 1999. 150(4): p. 281-285.
18. Casado López, S., et al., *Induction of genes encoding plant cell wall-degrading carbohydrate-active enzymes by lignocellulose-derived monosaccharides and cellobiose in the white-rot fungus Dichomitus squalens.* **Applied and Environmental Microbiology**, 2018. 84(11): p. e00403-18.
19. Peng, M., et al., *CreA-mediated repression of gene expression occurs at low monosaccharide levels during fungal plant biomass conversion in a time and substrate dependent manner.* **The Cell Surface**, 2021. 7: p. 100050.
20. de Vries, R.P., et al., *Expression profiling of pectinolytic genes from Aspergillus niger.* **FEBS Letters**, 2002. 530(1-3): p. 41-47.
21. Martens-Uzunova, E.S. and P.J. Schaap, *Assessment of the pectin degrading enzyme network of Aspergillus niger by functional genomics.* **Fungal Genetics and Biology**, 2009. 46(1): p. S170-S179.
22. Wu, V.W., et al., *The regulatory and transcriptional landscape associated with carbon utilization in a filamentous fungus.* **Proceedings of the National Academy of Sciences**, 2020. 117(11): p. 6003-6013.
23. Martinez, D., et al., *Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina).* **Nature Biotechnology**, 2008. 26(5): p. 553-560.
24. Li, J., et al., *The sugar metabolic model of Aspergillus niger can only be reliably transferred to fungi of its phylum.* **Journal of Fungi**, 2022. 8(12): p. 1315.
25. Gruben, B.S., et al., *Expression-based clustering of CAZyme-encoding genes of*

*Aspergillus niger.* **BMC Genomics**, 2017. 18(1): p. 1-18.

26. Filiatrault-Chastel, C., et al., *AA16, a new lytic polysaccharide monooxygenase family identified in fungal secretomes.* **Biotechnology for Biofuels**, 2019. 12(1): p. 1-15.
27. Li, L., C.J. Stoeckert, and D.S. Roos, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* **Genome Research**, 2003. 13(9): p. 2178-2189.
28. Emms, D.M. and S. Kelly, *OrthoFinder: phylogenetic orthology inference for comparative genomics.* **Genome Biology**, 2019. 20(1): p. 1-14.
29. Katoh, K., et al., *MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.* **Nucleic Acids Research**, 2002. 30(14): p. 3059-3066.
30. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.* **Molecular Biology and Evolution**, 2013. 30(4): p. 772-780.
31. Letunic, I. and P. Bork, *Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation.* **Nucleic Acids Research**, 2021. 49(W1): p. W293-W296.
32. de Vries, R.P., et al., *A new black Aspergillus species, A. vadensis, is a promising host for homologous and heterologous protein production.* **Applied and Environmental Microbiology**, 2004. 70(7): p. 3954-3959.
33. Klaubauf, S., et al., *Similar is not the same: differences in the function of the (hemi-) cellulolytic regulator XlnR (Xlr1/Xyr1) in filamentous fungi.* **Fungal Genetics and Biology**, 2014. 72: p. 73-81.
34. Eastwood, D.C., et al., *The plant cell wall–decomposing machinery underlies the functional diversity of forest fungi.* **Science**, 2011. 333(6043): p. 762-765.
35. Kryuchkova-Mostacci, N. and M. Robinson-Rechavi, *A benchmark of gene expression tissue-specificity metrics.* **Briefings in Bioinformatics**, 2017. 18(2): p. 205-214.
36. Yanai, I., et al., *Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification.* **Bioinformatics**, 2005. 21(5): p. 650-659.
37. van den Brink, J. and R.P. de Vries, *Fungal enzyme sets for plant polysaccharide degradation.* **Applied Microbiology and Biotechnology**, 2011. 91(6): p. 1477-1492.
38. Bornscheuer, U., K. Buchholz, and J. Seibel, *Enzymatic degradation of (ligno) cellulose.* **Angewandte Chemie International Edition**, 2014. 53(41): p. 10876-10893.
39. Ezeilo, U.R., et al., *Enzymatic breakdown of lignocellulosic biomass: the role of glycosyl hydrolases and lytic polysaccharide monooxygenases.* **Biotechnology &**

**Biotechnological Equipment**, 2017. 31(4): p. 647-662.

40. Kracher, D. and R. Ludwig, *Cellobiose dehydrogenase: An essential enzyme for lignocellulose degradation in nature–A review/Cellobiosedehydrogenase: Ein essentielles Enzym für den Lignozelluloseabbau in der Natur–Eine Übersicht.* **Die Bodenkultur: Journal of Land Management, Food and Environment**, 2016. 67(3): p. 145-163.
41. Langston, J.A., et al., *Oxidoreductive cellulose depolymerization by the enzymes cellobiose dehydrogenase and glycoside hydrolase 61.* **Applied and Environmental Microbiology**, 2011. 77(19): p. 7007-7015.
42. Collins, T., C. Gerday, and G. Feller, *Xylanases, xylanase families and extremophilic xylanases.* **FEMS Microbiology Reviews**, 2005. 29(1): p. 3-23.
43. Venegas, F.A., et al., *Carbohydrate esterase family 16 contains fungal hemicellulose acetyl esterases (HAEs) with varying specificity.* **New Biotechnology**, 2022. 70: p. 28-38.
44. Schultink, A., et al., *Structural diversity and function of xyloglucan sidechain substituents.* **Plants**, 2014. 3(4): p. 526-542.
45. Coutinho, P.M., et al., *Post-genomic insights into the plant polysaccharide degradation potential of Aspergillus nidulans and comparison to Aspergillus niger and Aspergillus oryzae.* **Fungal Genetics and Biology**, 2009. 46(1): p. S161-S169.
46. Damásio, A.R., et al., *Functional characterization and oligomerization of a recombinant xyloglucan-specific endo-β-1, 4-glucanase (GH12) from Aspergillus niveus.* **Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics**, 2012. 1824(3): p. 461-467.
47. Sun, P., et al., *Fungal glycoside hydrolase family 44 xyloglucanases are restricted to the phylum Basidiomycota and show a distinct xyloglucan cleavage pattern.* **Iscience**, 2022. 25(1): p. 103666.
48. Matsuzawa, T., et al., *Identification and characterization of two xyloglucan-specific endo-1, 4-glucanases in Aspergillus oryzae.* **Applied Microbiology and Biotechnology**, 2020. 104(20): p. 8761-8773.
49. Ademark, P., et al., *Cloning and characterization of Aspergillus niger genes encoding an α-galactosidase and a β-mannosidase involved in galactomannan degradation.* **European Journal of Biochemistry**, 2001. 268(10): p. 2982-2990.
50. Caffall, K.H. and D. Mohnen, *The structure, function, and biosynthesis of plant cell wall pectic polysaccharides.* **Carbohydrate Research**, 2009. 344(14): p. 1879-1900.
51. Benoit, I., et al., *Degradation of different pectins by fungi: correlations and contrasts between the pectinolytic enzyme sets identified in genomes and the growth on pectins of different origin.* **BMC Genomics**, 2012. 13(1): p. 1-11.
52. Mølgaard, A., S. Kauppinen, and S. Larsen, *Rhamnogalacturonan acetylesterase*

*elucidates the structure and function of a new family of hydrolases.* **Structure**, 2000. 8(4): p. 373-383.

53. Mischnick, P. and D. Momcilovic, *Chemical structure analysis of starch and cellulose derivatives.* **Advances in Carbohydrate Chemistry and Biochemistry**, 2010. 64: p. 117-210.
54. Taggart, P., *Starch as an ingredient: manufacture and applications.* **Starch in food: Structure, Function and Applications**, 2004: p. 363-392.
55. Gamage, A., et al., *Applications of starch biopolymers for a sustainable modern agriculture.* **Sustainability**, 2022. 14(10): p. 6085.
56. Roberfroid, M.B., *Introducing inulin-type fructans.* **British Journal of Nutrition**, 2005. 93(S1): p. S13-S25.
57. Pavis, N., et al., *Structure of fructans in roots and leaf tissues of Lolium perenne.* **New Phytologist**, 2001. 150(1): p. 83-95.
58. Zeuner, B., et al., *Comparative characterization of Aspergillus pectin lyases by discriminative substrate degradation profiling.* **Frontiers in Bioengineering and Biotechnology**, 2020. 8: p. 873.
59. Kun, R.S., et al., *The transcriptional activator ClrB is crucial for the degradation of soybean hulls and guar gum in Aspergillus niger.* **Fungal Genetics and Biology**, 2023. 165: p. 103781.
60. Yuan, X.-L., et al., *Identification of InuR, a new Zn (II) 2Cys6 transcriptional activator involved in the regulation of inulinolytic genes in Aspergillus niger.* **Molecular Genetics and Genomics**, 2008. 279: p. 11-26.
61. van Peij, N.N., et al., *The transcriptional activator XlnR regulates both xylanolytic and endoglucanase gene expression in Aspergillus niger.* **Applied and Environmental Microbiology**, 1998. 64(10): p. 3615-3619.
62. Gielkens, M.M., et al., *Two cellobiohydrolase-encoding genes from Aspergillus niger require D-xylose and the xylanolytic transcriptional activator XlnR for their expression.* **Applied and Environmental Microbiology**, 1999. 65(10): p. 4340-4345.
63. Sun, P., et al., *AA16 oxidoreductases boost cellulose-active AA9 lytic polysaccharide monooxygenases from Myceliophthora thermophila.* **ACS Catalysis**, 2023. 13(7): p. 4454-4467.
64. Agger, J.W., et al., *A new functional classification of glucuronoyl esterases by peptide pattern recognition.* **Frontiers in Microbiology**, 2017. 8: p. 309.
65. Ndeh, D., et al., *Complex pectin metabolism by gut bacteria reveals novel catalytic functions.* **Nature**, 2017. 544(7648): p. 65-70.
66. Suzuki, H., et al., *Biochemical characterization of a pectate lyase AnPL9 from Aspergillus nidulans.* **Applied Biochemistry and Biotechnology**, 2022. 194(12): p. 5627-5643.

67. Wilkens, C., et al., *GH62 arabinofuranosidases: structure, function and applications.* **Biotechnology Advances**, 2017. 35(6): p. 792-804.

68. Coconi Linares, N., et al., *Comparative characterization of nine novel GH51, GH54 and GH62 α-L-arabinofuranosidases from Penicillium subrubescens.* **FEBS Letters**, 2022. 596(3): p. 360-368.

## Supplementary materials

**Table S1.** CAZymes involved in plant polysaccharide degradation, and the number of putative CAZy-encoding genes in the six fungi related to the degradation of specific polysaccharides. ***Available upon request from the author***

**Table S2**. Orthologs of CAZymes detected among six studied species. ***Available upon request from the author***

**Table S3**. Summary of CAZy genes involved in plant polysaccharides degradation across the studied six fungi (including annotations from JGI). ***Available upon request from the author***

**Table S4**. Expression profiling of CAZymes encoding genes involved in plant polysaccharides degradation across the studied six fungi. ***Available upon request from the author***

**Tables S5**. Summary of CAZy genes with high sugar specificity in six fungi. ***Available upon request from the author***

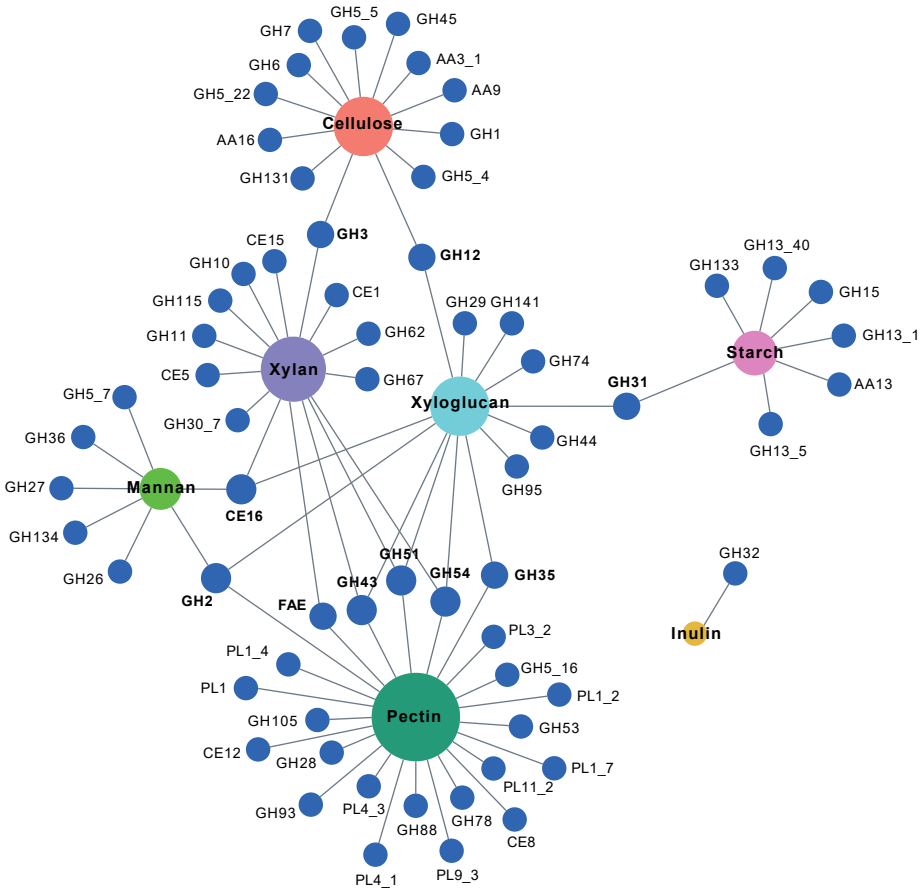


**Figure S1**. CAZy families related to plant biomass degradation. Families in bold are involved in the degradation of multiple polysaccharides.

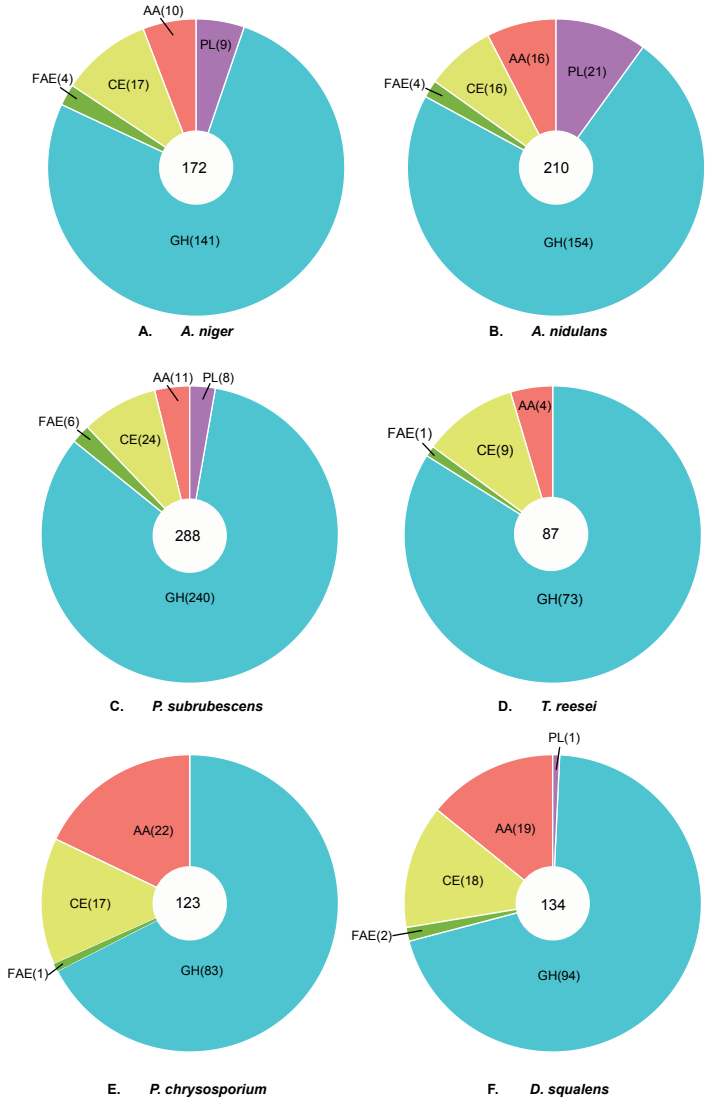


**Figure S2**. Pie-chart presenting the proportion of CAZy genes involved in plant biomass degradation in the genomes of the six fungi. GH: glycoside hydrolase, AA: Auxiliary activities, CE: Carbohydrate esterases, PL: Polysaccharide lyases, FAE: Feruloyl esterases.

**Figure S3**. Heatmap visualization of expression profiles of CAZy genes with sugar specific induction in the six fungi during their growth on nine monosaccharides. ***Available upon request from the author***

# Chapter 5

# Comparative transcriptomics reveals a diverse approach for the degradation of plant biomass of five filamentous fungi

Manuscript in preparation for publication

Mao Peng[#], Jiajia Li[#], Li Xu, Igor V. Grigoriev, Ronald P. de Vries

#These authors contributed equally

## Abstract

Fungal plant biomass conversion is of great importance in the Earth's carbon cycle, as well as in many industrial applications, such as the production of biofuel and biochemicals from plant lignocellulose materials. To obtain a deeper insight into the diversity of approaches that fungi employ to convert crude plant biomass, we comparatively analyzed the transcriptome profile of five diverse fungi during their growth on two common feedstocks, soybean hulls and corn stover, at three different time points. The selected fungi included the ascomycetes *Aspergillus niger*, *Aspergillus nidulans*, *Penicillium subrubescens*, *Trichoderma reesei* and the white-rot basidiomycete *Phanerochaete chrysosporium*. The gene expression profile of lignocellulose degrading enzymes, sugar transporters and sugar metabolic enzymes showed strong time-, substrate- and species-specific differences. By combining these transcriptome profiles on crude plant biomass to those on a diverse set of plant-derived monosaccharides, we were able to reveal complex gene regulation and substrate preference during the adaptation of these fungi to crude plant biomass. These results improve our understanding of the diversity and molecular mechanisms of plant biomass conversion by fungi, which will facilitate future studies of fungal ecological roles and further development of fungal biotechnology.

## 1. Introduction

Plant biomass is the most abundant and renewable carbon source in the world. The common polymers in plant biomass include cellulose, hemicellulose, pectin, lignin, as well as protein and storage polysaccharides (e.g., starch, gums and inulin). In nature, many fungi have evolved a sophisticated system for the efficient degradation and utilization of plant polymers. During the past decades, fungal plant biomass conversion (FPBC) has attracted increasing research interest due to its important ecological role in the Earth's carbon cycle, and huge industrial potential to produce biofuel, biochemicals and enzymes from renewable plant biomass [1].

FPBC is a highly dynamic and complex process which involves several crucial sets of genes. Firstly, a diverse set of fungal extracellular enzymes are required for degradation of plant biomass. Most of these enzymes belong to the carbohydrate-active enzymes (CAZymes) that has been well cataloged in the CAZy database (http://www.cazy.org/) [2], including glycoside hydrolases (GHs), carbohydrate esterases (CEs), polysaccharide lyases (PLs) and lytic polysaccharide monooxygenases (LPMOs). After enzymatic decomposition, a diverse set of mono- and oligosaccharides are released from plant biomass, which are transported into the cell mediated by various sugar transporters [3]. In addition, most fungi harbor a robust sugar metabolic network [4-6] that enables them to convert plant derived sugars into desired molecules and generate energy to support their growth and reproduction. Previous studies have demonstrated that fungi tailor their production of CAZymes to the plant substrate on which they grow [7-9]. In addition, analysis of fungal genomes has revealed that the content of lignocellulolytic CAZy genes varied significantly across fungal species [10-12]. Comparative functional genomics (e.g. transcriptome and proteome) have demonstrated even more significant diversity with respect to the activation and production of plant biomass degrading enzymes [13-17]. In contrast to the extensive studies of CAZymes, the comprehensive analysis of the (post-)genomics diversity

of sugar transporter and sugar metabolic networks severely lags behind. Recently, a limited number of studies have indicated the sugar transporters and sugar metabolic genes also showed a considerable diversity at genomic or post-genomics levels across different fungi [18, 19].

In this project, we aim to systematically evaluate the diversity of the fungal transcriptome response to lignocellulose, with the focus on expression profiles of CAZy enzymes, sugar transporters and carbon metabolic enzymes in five taxonomically diverse fungi grown on two common agriculture plant residues, corn stover (CS) and soybean hulls (SBH).

## 2. Materials and Methods

### *2.1 Strains, media and growth conditions*

The same set of five taxonomically diverse fungi were used that were also analysed in a previous study [18]. These include three Eurotiomycetes (*Aspergillus niger*, *Aspergillus nidulans*, *Penicillium subrubescens*), one Sordariomycete (*Trichoderma reesei*) and one Basidomycete (*Phanerochaete chrysosporium*).

The fungi were grown in minimal (MM) or complete (CM) medium as previously described [18]. In detail, *A. niger* N402, *A. nidulans* FGSC A4 and *P. subrubescens* CBS132785 were pre-cultured in complete medium [20] with 2% D-fructose. The same cultivation approach was used for *T. reesei* QM6a and *P. chrysosporium* PR-78, but with media optimized for these species [21, 22]. Mycelium was washed with MM and transferred to 250 ml Erlenmeyer flasks containing 50 ml MM supplemented with 1% CS or 1% SBH. Mycelium was harvested after cultivation for 4 h, 24 h and 48 h, by vacuum filtration, dried between tissue paper and frozen in liquid nitrogen.

### *2.2 RNA sequencing analysis*

Total RNA was extracted from ground mycelial samples using TRIzol reagent (Invitrogen, Breda, The Netherlands) according to the instructions

of the manufacturer. Purification of mRNA, synthesis of cDNA library and sequencing were conducted at DOE Joint Genome Institute (JGI), as previously described [23]. The replicates with poor reproducibility (Pearson correlation coefficient of raw counts < 0.9) were excluded for further statistical analysis.

DESeq2 v1.20 [24]t was used to determine differentially expressed genes (DEGs) between different growth conditions. Raw counts were used as DESeq2 input. The threshold of fold change (FC) ≥ 4 or ≤ 0.25 and adjusted P-value < 0.01 was used to define significant DEGs (higher-/lower- expressed genes). Besides the newly generated transcriptome data of fungi grown on two crude plant biomasses (CS and SBH), we also integrated and analyzed transcriptome data of fungi grown on nine different monosaccharides from our previous study [18]. Carbon metabolic enzymes and pathways were extracted from our previous study [18]. The CAZy family information was obtained from JGI MycoCosm database, and annotation of plant polysaccharide specificity of CAZymes was based on previously publications [9, 17]. Heatmap and clustering were plotted using R package "ComplexHeatmap" [25]. The principal component analysis (PCA) was performed with the DEseq2 using the normalized counts by applying the '*regularized log (rlog)*' method. Gene set variation analysis (GSVA) [26] was performed to estimate variation of pathway activity over the tested conditions based on the average values of gene expression (fragments per kilobase of transcript per million fragments mapped, FPKMs) in each studied condition, and the analysis mainly focused on the genes encoding preselected CAZy and sugar metabolic enzymes. The RNA-seq data generated in this study is available at the JGI portal (proposal ID 504291).



## 3. Results and discussion

### *3.1 Transcriptome analysis revealed strong condition-dependent and species-specific responses*

We performed comparative transcriptome analysis of five fungi to study their molecular response to different plant biomass at different time points. The

principal component analysis (Figure 1) and differentially expressed genes (DEGs) analysis (Figure 2 and Table S1) revealed that fungal adaption to plant biomass showed strong time-, substrate- and species-specific patterns. We compared the expression on the two crude plant biomass substrates to expression on D-glucose, as on this substrate typically CAZy genes are not or poorly expressed. All studied fungi grown on CS and SBH showed significantly different transcriptome profiles in terms of the number of genes showing differential expression relative to D-glucose. Moreover, a considerable diversity of their response to CS and SBH was observed, indicating that fungi tailor their transcriptome response to the composition of the plant biomass substrate. The transcriptome of *A. niger* on CS at different time points showed only a minimal difference, while its transcriptome on SBH showed a clear time-specific response. For *A. nidulans*, large differences in its response to the two tested substrates was identified at the early time point (4 h), while very similar expression profiles were observed at the later time points (24 h and 48 h) (Figure 1). *P. subrubescens* displayed the strongest substrate- and time-specific response, with a relatively large set of differentially expressed genes that were identified in each comparison and clearly separated substrate-time course trajectories in the PCA. *T. reesei* showed a similar expression behavior as *A. nidulans*, with many significantly different expression profiles at the early time point (4 h) and only small sets of DEGs detected at later time points for both types of plant biomass substrates. *P. chrysosporium* presented a minimal substrate specific response at 4 h compared to the other four species, in which only 68 DEGs were detected. However, a clear differential response was identified between 4 h and the later time points (24 h or 48 h) for both plant biomass substrates. These differences in expression patterns most likely reflect a different organization of the regulatory system controlling these genes in the studied fungi.

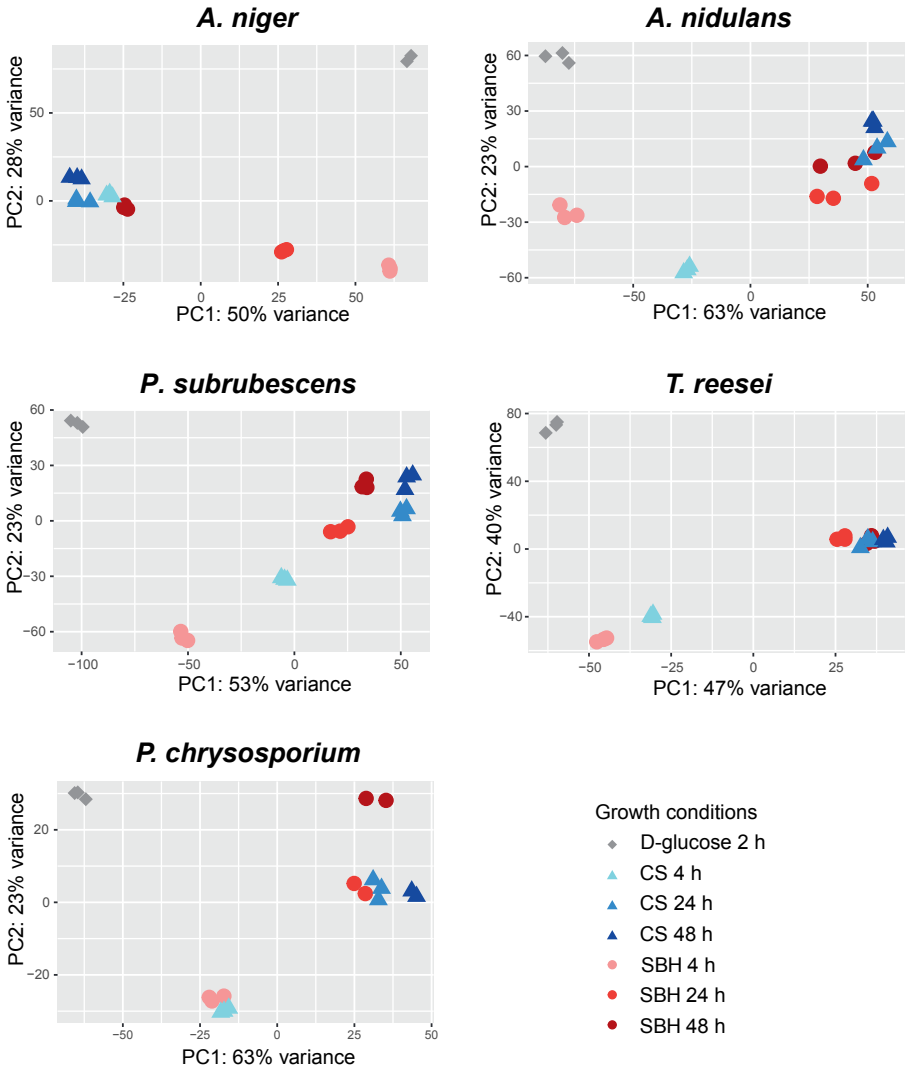


**Figure 1. Principal component analysis (PCA) of transcriptome data**. PCA was performed on normalized counts of the whole transcriptome. Samples grown on different carbon sources (i.e., glucose, corn stover (CS) and soybean hulls (SBH)) at different time points are indicated with different shapes and colors.

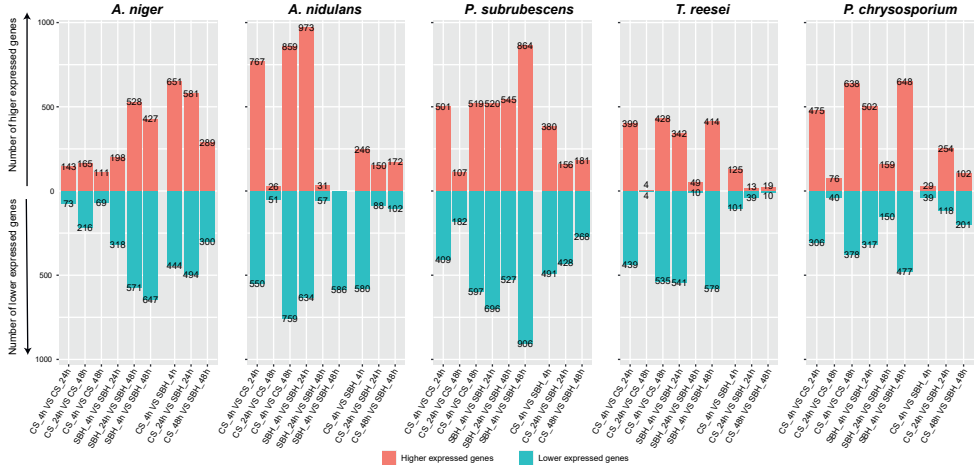


**Figure 2. The total number of differentially expressed genes identified in comparison of fungi grown on corn stover (CS) and soybean hulls (SBH) at three time points (4 h, 24 h and 48 h).** The total number of higher and lower expressed genes in each comparison were indicated in red and blue, respectively.

## *3.2 Expression diversity of key genes related to plant biomass conversion across the fungi*

Several important biological processes and related genes have been well-characterized to play essential role during FPBC. These include the CAZymes involved in lignocellulose degradation, sugar transporters mediating transmembrane uptake of small sugars, sugar catabolic enzymes for intracellular metabolism of sugars and transcription factors that control the whole system [27]. We therefore analyzed the expression profiles of these four categories of genes in more detail.

### *3.2.1 Expression profile of CAZy genes related to plant biomass degradation (PBD)*

Our previous study revealed that taxonomically close fungi often have a similar CAZy content (see **Chapter 4** of this thesis). However, our transcriptome analysis revealed that a similar genomic profile of FPBC related CAZy genes did not guarantee a similar transcriptomic profile (Figure 3 and Figure S1).

For instance, the genomes of *A. niger* and *A. nidulans* have a comparable set of CAZy genes involved in FPBC (Table S2), but these genes show distinct transcriptome profiles, indicating the influence of the differences in the regulatory system that controls their expression.

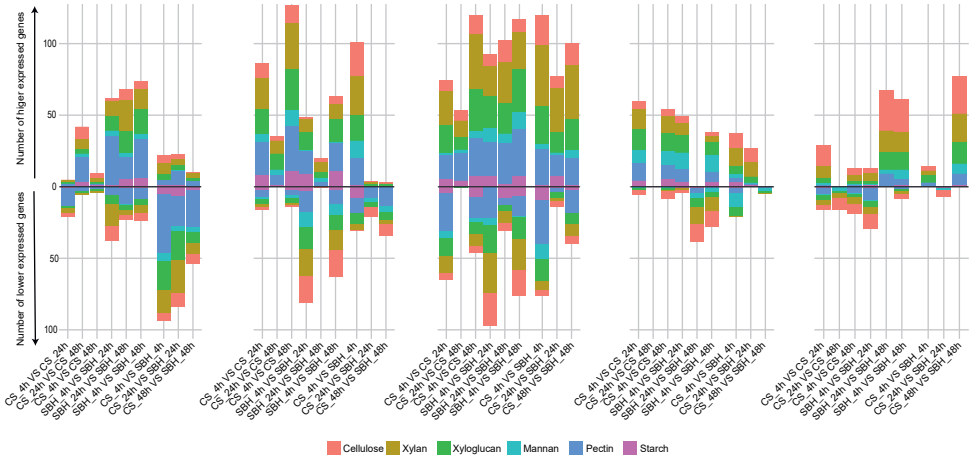


**Figure 3. Differentially expressed FPBC related CAZy genes when comparing fungi grown on corn stover (CS) and soybean hulls (SBH) at three time points (4 h, 24 h and 48 h).** The number of genes involved in decomposition of different plant polymers are indicated in different colors. The bars represent the number of differentially expressed genes between two conditions. The upper bars show genes that are higher expressed in the first condition for each comparison, while the lower bars show genes that are higher expressed in the second condition.

The expression profile of these CAZy genes during growth of *A. niger* on CS at different time points showed much less difference compared to its profile on SBH (Figure 3). A considerable number of CAZymes genes involved in xylan, xyloglucan, pectin and starch degradation were more actively expressed on SBH than on CS, while specific temporal changes were also observed during its growth on SBH. Notably, a high number of pectinolytic genes of *A. niger* were significantly induced at 4 h compared to 24 h or 48 h during fungus grown on SBH. Although both the CS and SBH contain abundant cellulose, a relatively small number of cellulolytic genes were differentially expressed during the growth of *A. niger* on these substrates compared to several of the other studied fungi (except *T. reesei*) (Figure 3).

For *A. nidulans*, many FPBC related CAZy genes showed higher expression at the early time point (4 h) on both CS and SBH, and much less CAZy genes were identified as DEGs in the comparison between 24 h and 48 h. In *A. nidulans*, CS induced a higher number of genes related to degradation of cellulose, xylan, xyloglucan and mannan at 4 h, while SBH specifically induced starch degrading genes at 4 h and induced some genes involved in cellulose, mannan and pectin degradation at 24 h and 48 h.

*P. subrubescens* continuously tailored its expression of FPBC genes during growth on both substrates. The genes involved in degradation of cellulose, xylan, xyloglucan, pectin, and starch showed strong temporal expression profiles on both substrates. The comparison between CS and SBH showed that CS induced more genes related to degradation of cellulose, xylan and xyloglucan at all time points, while SBH induced more genes involved in mannan degradation at 4 h.

*T. reesei* has a relatively small size of genes involved in FPBC [28] (see **Chapter 4** of this thesis). Growth of *T. reesei* on different conditions identified less differentially expressed FPBC genes compared to the three Eurotiomycetes. The FPBC genes of *T. reesei* were predominantly induced at the early time point (4 h) during its growth on CS, while some genes involved in cellulose, xylan and mannan were induced at 48 h compared to 4 h or 24 h on SBH. Compared to growth on SBH, growth on CS induced more genes at 4 h and 24 h related to degradation of cellulose and xylan, while some mannanolytic genes were specifically induced on SBH at 4 h.

For *P. chrysosporium*, we also identified a relatively small set of FPBC related genes that were differentially expressed. During growth of the fungus on CS, higher expression of cellulololytic genes was observed at 4 h and 48 h compared to 24 h, while xyloglucanolytic genes were mainly induced at 4 h. In contrast, during growth of the fungus on SBH, most FPBC genes showed low expression at 48 h (Figure S1). The comparison between CS and SBH identified some FPBC genes at 4 h and 24 h, but a considerable number of

genes related to cellulose, xylan and xyloglucan genes showed significantly higher expression on CS than SBH at 48 h.

Overall, the diverse expression profiles of FPBC related CAZy genes are in line with the polysaccharide composition of two tested plant biomass substrates and polysaccharide preference of the tested fungal species. For instance, the higher abundance of xylan in CS than in SBH induced a larger number of xylan degrading genes during fungal growth on CS than on SBH for most of the studied fungi (except *A. niger*). Compared to other species, a relatively small set of cellulolytic and pectinolytic genes showed differential expression, respectively, during growth of *A. niger* and *P. chrysosporium*, respectively, which is in line with the relatively poor growth of *A. niger* on cellulose, and less efficient use of pectin for *P. chrysosporium* (see **Chapter 4** of this thesis). This demonstrates that both the genome content as well as the expression profiles correlate well with growth for these two species.

*3.2.2 Expression profile of genes related to sugar metabolism (SM)*

The comparison of expression profile of SM genes also revealed clear diversity across different studied fungi (Figure 4 and Figure S2). For *A. niger*, a higher number of SM related genes were highly expressed and their expression changed dynamically during growth on SBH compared to CS. The genes involved in metabolism of D-galactose, D-galacturonic acid, pentose catabolic pathway (PCP) and tricarboxylic acid (TCA) cycle were the major temporally changed genes during growth on SBH.

For *A. nidulans*, similar as for the FPBC related CAZy genes, a larger number of differentially expressed SM genes were identified when comparing early (4 h) and later time points (24 h or 48 h) than between the two later time points. The comparison of time points on both plant substrates showed that SM genes related to the D-galacturonic acid pathway and the PCP were predominantly induced at 4 h, while genes related to glycolysis and glycerol metabolism were mainly induced at 24 h and 48 h. The comparison of *A. nidulans* grown

on CS and SBH showed that SM genes related to the D-galacturonic acid pathway were higher expressed on SBH at all tested time points, which is consistent with the corresponding sugar composition of these two substrates [29]. In addition, a few genes related to D-galactose metabolism were higher expressed on SBH at 24 h and 48 h than CS.

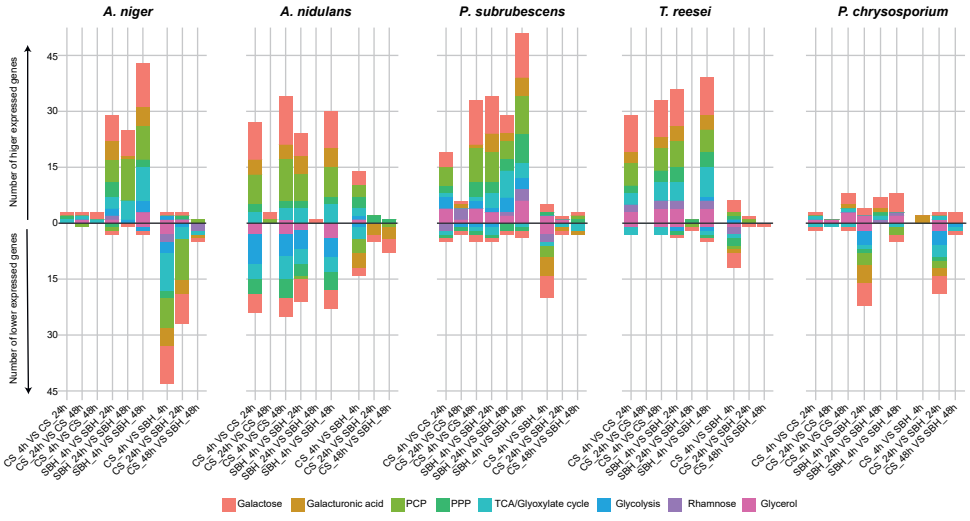


**Figure 4. Differentially expressed sugar metabolic genes when comparing fungi grown on corn stover (CS) and soybean hulls (SBH) at three time points (4 h, 24 h and 48 h).** The number of genes involved in different sugar metabolic pathways are indicated in different colors.

The comparison of *P. subrubescens* showed a slightly different expression profile compared to the trend of the CAZy profile discussed above. For instance, the majority of SM genes were induced at the early time point, while we observed a comparable number of CAZy genes that were specifically induced at both early and later time points. The most obvious expression difference between *P. subrubescens* grown on CS and SBH occurred at 4 h, where a significant higher number of SM genes involved in the metabolism of D-galactose, D-galacturonic acid, PCP and D-rhamnose showed higher expression on SBH than on CS.

The SM genes of *T. reesei* during growth on both plant substrates showed a

similar expression profile as observed for *P. subrubescens*, in which most genes were predominantly induced at 4 h and only few genes were differentially expressed between 24 h and 48 h.

Compared to the other species, a relatively small set of SM genes of *P. chrysosporium* showed significant differential expression, except that during growth on SBH at 24 h a significant larger number of SM genes was induced than in the other conditions (Figure 4 and Figure S2). In contrast, the most significant transcriptome change of FPBC related CAZy genes was identified at 48 h during growth *P. chrysosporium* on CS and SBH (Figure 3).

*3.2.3 Expression profile of sugar transporters (STs)*

The genomes of here studied fungi encode variable numbers of sugar transporters (STs). Based on a computational search for conserved protein domains that define STs, we identified 90, 83, 117, 52 and 23 STs for *A. niger*, *A. nidulans*, *P. subrubescens*, *T. reesei* and *P. chrysosporium*, respectively (Table S2). The expression of these STs showed generally similar expression trends as CAZy genes and SM genes. We identified distinct differences in the expression of STs during growth of *A. niger* on SBH and CS. A minimal set of STs showed expression changes during growth of *A. nidulans* and *T. reesei* on plant substrates at the two later time points (24 h and 48 h). *P. subrubescens* tailored the expression of a specific set of STs for adaptation to each tested condition, while only few STs of *P. chrysosporium* showed significant differential expression (Figure 5).

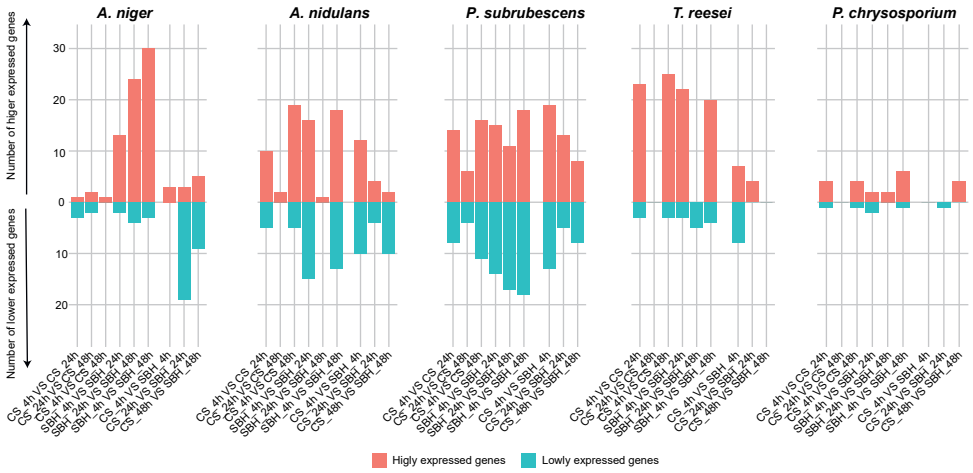


**Figure 5. Differentially expressed sugar transport genes when comparing fungi grown on corn stover (CS) and soybean hulls (SBH) at three time points (4 h, 24 h and 48 h)**. The total number of higher and lower expressed genes in each comparison were indicated in red and blue, respectively.

### *3.2.4 Diverse expression profile of transcription factors (TFs)*

The dynamic change of gene expression is driven by TFs. Previous studies have identified a set of key TFs related to FPBC and suggested their functional diversity across different species [30], even though only three TFs for FPBC have so far been reported in Basidiomycetes (i.e., *cre1* [31], *pacC* [32] and *ace3*/*roc1* [30, 33]). Comparing the expression profile of these TFs (both experimentally characterized and ortholog-based predicted TFs) across our tested species revealed a strong transcriptional diversity for most of the selected TFs (Figure 6). For example, the highly conserved TFs *creA/cre1* of *A. niger* showed higher expression on SBH than CS at 4 h, while its ortholog in *A. nidulans* had higher expression on CS than SBH at 4 h and 24 h, and higher expression in *P. subrubescens* occurred on both SBH and CS at 4 h. In contrast, the overall expression of *creA/cre1* orthologs in *T. reesei* and *P. chrysosporium* was low in all tested conditions.

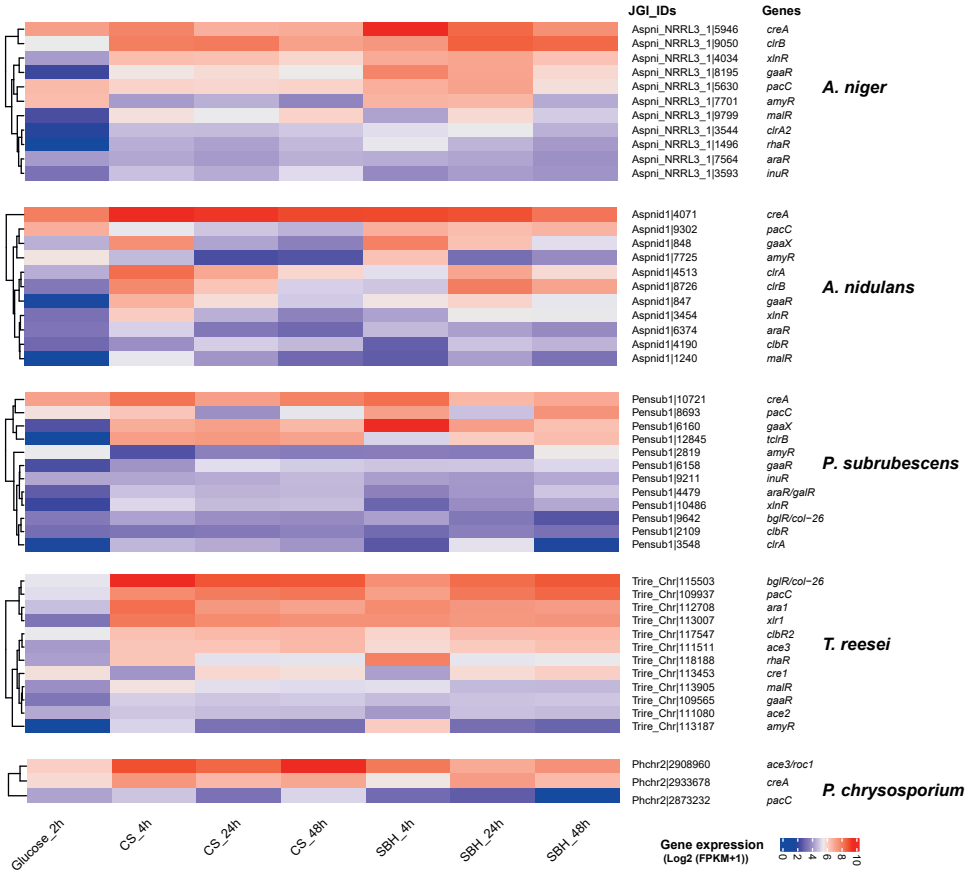


**Figure 6. Gene expression profiles of known and predicted transcription factors involved in FPBC.**

The pH-responsive TF, *pacC,* is another conserved regulator that has been suggested to be involved in FPBC. Expression of *pacC* was higher during growth on SBH than on CS for the three Eurotiomycetes, while it showed overall high and low expression for *T. reesei* and *P. chrysosporium*, respectively, on both tested plant substrates.

In addition, the expression profile of a crucial TF involved in pectin utilization, *gaaR*, showed higher expression on SBH at 4 h and 24 h for *A. niger*, and higher expression on CS at 4 h for *A. nidulans*, while no clear differential expression was identified for the other Ascomycetes.

The genes encoding the starch utilization regulator AmyR showed higher expression on SBH at 4 h for *A. niger*, *A. nidulans* and *T. reesei*, while its overall expression was low in *P. subrubescens*.

The *ace3* gene has been characterized as an important TF for regulating cellulose degradation in *T. reesei* [34] and Basidiomycete fungi [30, 33]. In our expression profiles, this gene showed relative high expression in *T. reesei* and *P. chrysosporium* grown on both plant substrates. The highest expression of ace3 during growth of *P. chrysosporium* on CS at 48 h is consistent with the general higher expression of cellulase genes in this condition.

The strong species-specific expression profile of TFs shown here already indicates a complex and fast evolving regulatory networks govern the diversity of FPBC process. In addition, the post-transcriptional activation of transcription factors will further affect the regulatory system in each species, likely contributing also strongly to the differences in expression profiles observed between the studied fungi.

### *3.3 Transcriptomic comparison of plant biomass substrates and monosaccharides provides insights into the polysaccharide preference of some of these fungi*

To reveal the sugar preference and interaction between FPBC-related CAZy and SM genes during fungal utilization of complex plant polymers, we combined and analyzed transcriptome data of fungi grown on both crude plant biomass substrates and nine monosaccharides. GSVA analysis [26] was applied to estimate the variation of gene expression levels over the test conditions, which revealed a considerable difference across different samples and fungal species (Figure 7). As expected, the enrichment profiles obtained from the GSVA analysis on crude plant biomass samples are consistent with the DEGs analysis results discussed above. For instance, overall higher expression levels of genes encoding CAZymes and sugar metabolic enzymes were identified on SBH at 4 h and 24 h for *A. niger*, on both CS and SBH at 4 h for *T. reesei*, and on SBH at 24 h for *P. chrysosporium*.

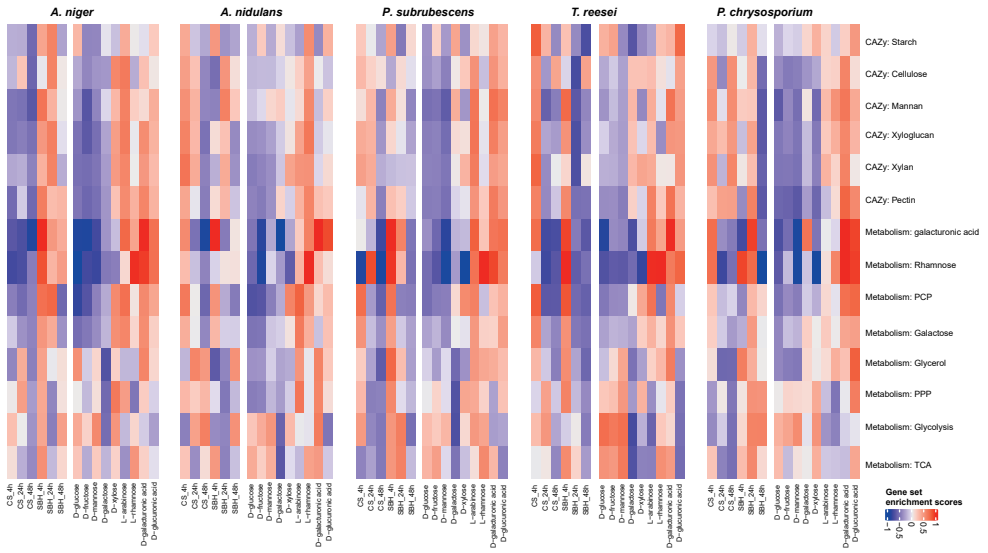


**Figure 7. Gene set variation analysis of FPBC related CAZy and SM genes across transcriptome data of five fungi grown on two crude plant biomass and nine different monosaccharides.** The x-axis shows the different carbon sources used in the cultures. CS and SBH were analyzed at three time points, while all monomeric carbon sources were only analyzed after 2 h.

The detailed expression levels of CAZy genes during growth on monosaccharides showed huge differences across species. Many cellulolytic and hemicellulolytic CAZy genes of *A. niger* showed higher levels during growth on two pentoses (L-arabinose and D-xylose), while the pectinolytic CAZy genes were induced on D-galacturonic acid and L-arabinose. The CAZy genes of *P. subrubescens* and *T. reesei* showed a similar expression profile as those of *A. niger*, except that D-xylose showed no clear induction for CAZy genes of *P. subrubescens* and the two pentose sugars showed relative weak induction for (hemi-)cellulolytic genes of *T. reesei*. In contrast, the expression of CAZy genes in *A. nidulans* and *P. chrysosporium* was mainly induced by L-rhamnose and D-galacturonic acid, respectively. These results confirm the strong influence of the differences in the regulatory system that control which enzymes are produced on each plant biomass substrate.

The expression profile of SM genes showed considerable similarity in

monosaccharide induction patterns among the species. For instance, the SM genes related to TCA and glycolysis in all five species showed relative high expression on easily metabolizable sugars, i.e., D-glucose, D-fructose and D-mannose. The genes related to metabolism of D-galacturonic acid, L-rhamnose and pentose showed as expected higher expression levels on D-galacturonic acid, L-rhamnose and the two pentoses, respectively. In contrast, the L-rhamnose and pentose metabolic genes of *P. chrysosporium* were not induced in the corresponding monosaccharide samples, but were slightly induced by D-galacturonic acid and D-glucuronic acid. In contrast, the expression profile of SM genes showed clear diversity across crude plant biomass samples among the species. The higher expression levels of each specific group of SM were not always identified on the same plant substrate at the same time points in different species.

The combined analysis of the expression profiles of CAZy and SM genes (Figure 7) together with the chemical composition of the two tested plant biomass substrates provided insights into the polysaccharide preference of these fungi during growth on complex plant biomass.

*A. niger* appears to respond poorly to CS, which matches its relatively poor growth on this substrate [29]. In contrast, the clear temporal expression profile on SBH reveals that its primary response is mainly aimed at pectin degradation and metabolism of D-galacturonic acid, L-rhamnose and pentose, the main sugar components of this polysaccharide. At 24 h, expression of genes related to xylan, xyloglucan, starch and cellulose are induced, suggesting that they are of secondary preference as a carbon source compared to pectin for *A. niger*. SBH contains a significant amount of pectin, but also the other types of polysaccharides, suggesting a clear pectin-favored approach for this species.

*A. nidulans*, *P. subrubescens* and *T. reesei* all have a broad response to CS at the early time point in which genes related to the degradation of all polysaccharide components are induced simultaneously. More differences are observed on SBH, where *A. nidulans* initially induces amylolytic genes, while after 24 h

genes related to the degradation of most other polysaccharides are induced. In contrast, *P. subrubescens* induces both amylolytic and mannanolytic genes at 4 h, with no clear induction of the other polysaccharide related genes at 24 h. Interestingly, both fungi induce genes of D-galacturonic acid metabolism (and also L-rhamnose metabolism for *P. subrubescens*) at 4 h despite no clear induction of pectinolytic genes. A possible explanation for this could be an even earlier (and shorter) induction of these genes before the 4 h sample was taken. *T. reesei* induced mannanolytic and pectinolytic genes at 4 h on SBH, together with genes of D-galacturonic acid metabolism, L-rhamnose metabolism and the PCP. *T. reesei* has a relatively small pectinolytic enzyme set, but appears to prefer this component over xylan and cellulose during growth on SBH.

The basidiomycete *P. chrysosporium* partially followed the pattern of *A. nidulans*, *P. subrubescens* and *T. reesei*, with a broad induction response to CS, but excluding induction of amylolytic genes. On SBH, clearest induction of polysaccharide degradation related genes occurred after 24 h with again a broad response, together with induction of most carbon metabolism related genes. The broad response of this fungus could be related to SBH and CS being non-natural substrates for this fungus as it naturally grows on woody substrates that have a significantly different composition [35]. *P. chrysosporium* shows a similar response to D-galacturonic acid and D-glucuronic acid. While the first compound is a main component of pectin which is only marginally present in wood, the second is a common side chain of wood xylans [36]. It would therefore be conceivable that the induction system in this species does not distinguish between these two uronic acids, although the presence of the genes of the D-galacturonic acid and L-rhamnose pathways also suggests that this metabolic ability has been maintained in *P. chrysosporium* even though they are not the main sugars in its natural habitat.

The differences between *A. nidulans* and *A. niger* on SBH mirror those observed for another dicot feedstock (sugar beet pulp), in which a proteomic

analysis revealed a higher production of cellulases for *A. nidulans*, while *A. niger* produced largely pectinases [37]. Interestingly, on this substrate *P. subrubescens* produced highly similar enzyme activities as *A. niger* [38], while in our study it was highly similar to *A. nidulans*, suggesting that culture conditions and sample times strongly affect these comparisons.

## 4. Conclusions

In this study, we comparatively analyzed the transcriptomes of five fungi grown on two crude plant biomasses. The gene expression profiles of lignocellulose degrading enzymes, sugar transporter and sugar metabolic genes showed strong time-, substrate- and species-specific manner, which highlights the diverse approaches of these fungal species for adaptation to crude plant materials and their complex transcription regulation. These results enhanced our understanding of ecological role of fungi in FPBC and will facilitate development of fungal biotechnology for better use of lignocellulose as a renewable resource. They also demonstrate that comparison purely based on genome content can provide a misleading picture, as the expression of the genes strongly affects the approach used by each fungus and not always follows the same pattern as the genome content.

## Author Contributions

All authors contributed to the manuscript. R.P.d.V. conceived and supervised the overall project. M.P., J.L. and L.X. performed the formal analysis and M.P. wrote the original draft. M.P. and R.P.d.V. revised the manuscript.

## Acknowledgements

JL was supported by the China Scholarship Council (CSC 201909110079).

## References

1. Meyer, V., et al., *Growing a circular economy with fungal biotechnology: a white paper.* **Fungal Biology and Biotechnology**, 2020. 7: p. 5.
2. Drula, E., et al., *The carbohydrate-active enzyme database: functions and literature.* **Nucleic Acids Research**, 2022. 50(D1): p. D571-D577.

3. Peng, M., et al., *In Silico analysis of putative sugar transporter genes in Aspergillus niger using phylogeny and comparative transcriptomics.* **Frontiers in Microbiology**, 2018. 9: p. 1045.
4. Aguilar-Pontes, M.V., et al., *The gold-standard genome of Aspergillus niger NRRL 3 enables a detailed view of the diversity of sugar catabolism in fungi.* **Studies in Mycology**, 2018. 91: p. 61-78.
5. Khosravi, C., et al., *Sugar catabolism in Aspergillus and other fungi related to the utilization of plant biomass.* **Advances in Applied Microbiology**, 2015. 90: p. 1-28.
6. Chroumpi, T., et al., *Revisiting a 'simple' fungal metabolic pathway reveals redundancy, complexity and diversity.* **Microbial Biotechnology**, 2021. 14: p. 2525-2537.
7. Dilokpimol, A., et al., *Penicillium subrubescens adapts its enzyme production to the composition of plant biomass.* **Bioresource Technology**, 2020. 311: p. 123477.
8. Wu, V.W., et al., *The regulatory and transcriptional landscape associated with carbon utilization in a filamentous fungus.* **Proceedings of the National Academy of Sciences**, 2020. 117(11): p. 6003-6013.
9. Daly, P., et al., *Dichomitus squalens partially tailors its molecular responses to the composition of solid wood.* **Environmental Microbiology**, 2018. 20(11): p. 4141-4156.
10. Zhao, Z., et al., *Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi.* **BMC Genomics**, 2013. 14: p. 274.
11. de Vries, R.P., et al., *Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus Aspergillus.* **Genome Biology**, 2017. 18(1): p. 28.
12. Riley, R., et al., *Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi.* **Proceedings of the National Academy of Sciences**, 2014. 111(27): p. 9923-8.
13. de Vries, R.P. and M.R. Makela, *Genomic and postgenomic diversity of fungal plant biomass degradation approaches.* **Trends in Microbiology**, 2020. 28(6): p. 487-499.
14. Arntzen, M.O., et al., *Quantitative comparison of the biomass-degrading enzyme repertoires of five filamentous fungi.* **Scientific Reports**, 2020. 10(1): p. 20267.
15. Kijpornyongpan, T., et al., *Systems biology-guided understanding of white-rot fungi for biotechnological applications: A review.* **iScience**, 2022. 25(7): p. 104640.
16. Vanden Wymelenberg, A., et al., *Significant alteration of gene expression in wood decay fungi Postia placenta and Phanerochaete chrysosporium by plant species.* **Applied and Environmental Microbiology**, 2011. 77(13): p. 4499-507.
17. Benoit, I., et al., *Closely related fungi employ diverse enzymatic strategies to degrade plant biomass.* **Biotechnology for Biofuels**, 2015. 8: p. 107.

18. Li, J., et al., *The sugar metabolic model of Aspergillus niger can only be reliably transferred to fungi of its phylum.* **Journal of Fungi**, 2022. 8(12).
19. Resl, P., et al., *Large differences in carbohydrate degradation and transport potential among lichen fungal symbionts.* **Nature Communications**, 2022. 13(1): p. 2634.
20. de Vries, R.P., et al., *A new black Aspergillus species, A. vadensis, is a promising host for homologous and heterologous protein production.* **Applied and Environmental Microbiology**, 2004. 70(7): p. 3954-9.
21. Klaubauf, S., et al., *Similar is not the same: differences in the function of the (hemi-) cellulolytic regulator XlnR (Xlr1/Xyr1) in filamentous fungi.* **Fungal Genetics and Biology**, 2014. 72: p. 73-81.
22. Eastwood, D.C., et al., *The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi.* **Science**, 2011. 333(6043): p. 762-5.
23. Chroumpi, T., et al., *Identification of a gene encoding the last step of the L-rhamnose catabolic pathway in Aspergillus niger revealed the inducer of the pathway regulator.* **Microbiological Research**, 2020. 234: p. 126426.
24. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* **Genome Biology**, 2014. 15(12): p. 550.
25. Gu, Z., R. Eils, and M. Schlesner, *Complex heatmaps reveal patterns and correlations in multidimensional genomic data.* **Bioinformatics**, 2016. 32(18): p. 2847-9.
26. Hanzelmann, S., R. Castelo, and J. Guinney, *GSVA: gene set variation analysis for microarray and RNA-seq data.* **BMC Bioinformatics**, 2013. 14: p. 7.
27. Peng, M., et al., *CreA-mediated repression of gene expression occurs at low monosaccharide levels during fungal plant biomass conversion in a time and substrate dependent manner.* **The Cell Surface**, 2021. 7: p. 100050.
28. Martinez, D., et al., *Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina).* **Nature Biotechnology**, 2008. 26(5): p. 553-60.
29. Khosravi, C., et al., *Transcriptome analysis of Aspergillus niger xlnR and xkiA mutants grown on corn Stover and soybean hulls reveals a highly complex regulatory network.* **BMC Genomics**, 2019. 20(1): p. 853.
30. Benocci, T., et al., *Regulators of plant biomass degradation in ascomycetous fungi.* **Biotechnology for Biofuels**, 2017. 10: p. 152.
31. Pareek, M., et al., *Preassembled Cas9 ribonucleoprotein-mediated gene deletion identifies the carbon catabolite repressor and its target genes in Coprinopsis cinerea.* **Applied and Industrial Microbiology**, 2022. 88(23): p. e0094022.
32. Zhu, J., et al., *Improvement of laccase activity by silencing PacC in Ganoderma lucidum.* **World Journal of Microbiology and Biotechnology**, 2022. 38(2): p. 32.
33. Marian, I.M., et al., *The transcription factor Roc1 is a key regulator of cellulose*

*degradation in the wood-decaying mushroom Schizophyllum commune.* **mBio**, 2022. 13(3): p. e0062822.

34. Zhang, J., et al., *The transcription factor ACE3 controls cellulase activities and lactose metabolism via two additional regulators in the fungus Trichoderma reesei.* **Journal of Biological Chemistry**, 2019. 294(48): p. 18435-18450.
35. Martinez, D., et al., *Genome sequence of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78.* **Nature Biotechnology**, 2004. 22(6): p. 695-700.
36. Curry, T.M., M.J. Pena, and B.R. Urbanowicz, *An update on xylan structure, biosynthesis, and potential commercial applications.* **The Cell Surface**, 2023. 9: p. 100101.
37. Makela, M.R., et al., *Genomic and exoproteomic diversity in plant biomass degradation approaches among Aspergilli.* **Studies in Mycology**, 2018. 91: p. 79-99.
38. Makela, M.R., et al., *Penicillium subrubescens is a promising alternative for Aspergillus niger in enzymatic plant biomass saccharification.* **New Biotechnology**, 2016. 33(6): p. 834-841.

## Chapter 5

# Supplementary materials

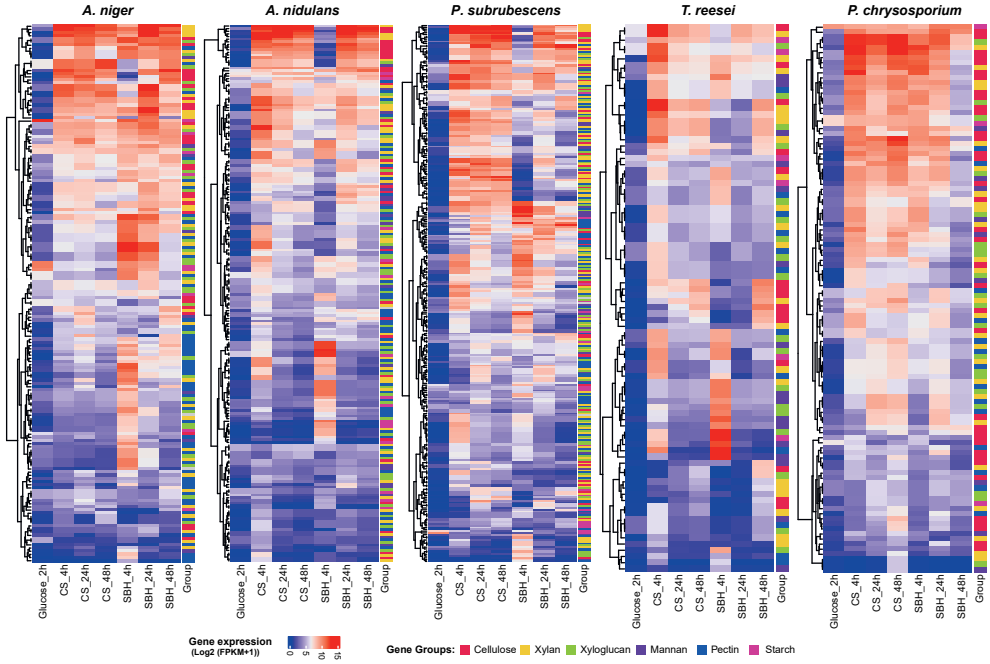


**Figure S1.** Expression profiles of CAZy genes involved in plant biomass conversion.

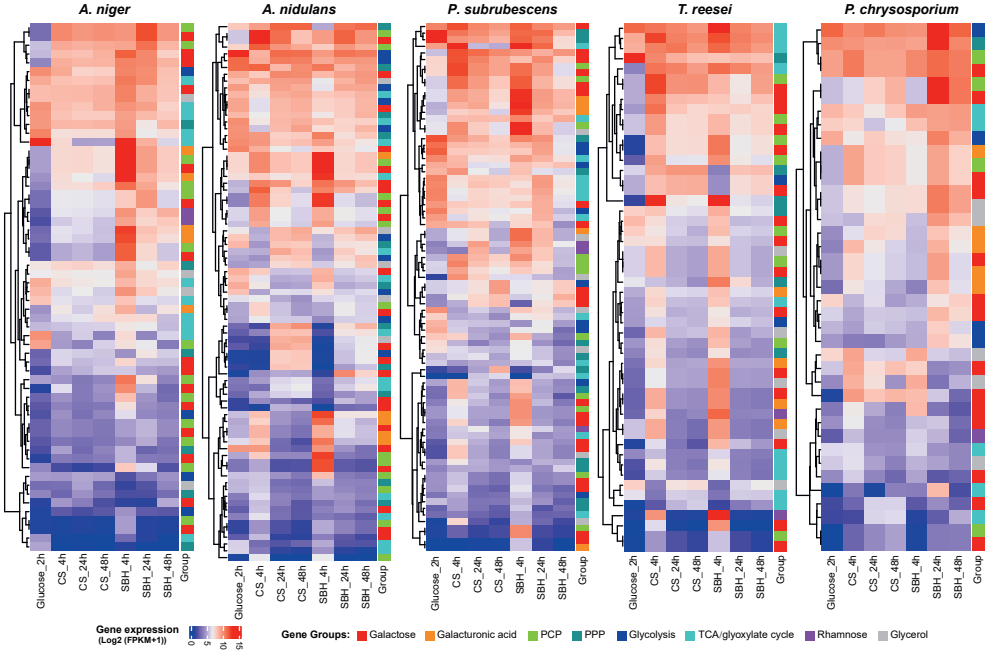


**Figure S2.** Expression profiles of sugar catabolic enzyme encoding genes.

**Table S1**. Gene expression levels and statistical analysis of all genes for five fungi grown in different conditions. ***Available upon request from the author***

**Table S2**. The list of genes predicted to be involved in plant biomass conversion for the five studied fungi. ***Available upon request from the author***

# Chapter 6

# Summary and general discussion

In this thesis, I focused on two important aspects of plant biomass conversion: primary sugar metabolism and largely secreted CAZymes involved in the degradation of complex plant biomass substrates. The main objective of this thesis was to improve our understanding of the diversity of primary carbon metabolism and the enzymatic ability of fungi. These insights are here discussed in a more general context and were also supplemented with additional results obtained during my studies that were not included in the previous chapters.

## 1. The diversity of sugar metabolism across the fungal kingdom

Knowledge on sugar catabolism is not only relevant to understand the roles and abilities of fungi in natural habitats, but also for their potential as synthetic pathways for the production of biochemicals. In **Chapter 3** [1], I describe the potential and limits of extrapolating the reference sugar metabolic model of *A. niger* to other fungi using an orthology-based approach combined with omics-analysis on relevant carbon sources to compare the multi-omics response in different fungi. The results revealed that reliable transfer of sugar metabolic pathways is only possible to the close related species. *A. nidulans* has a smaller taxonomic distance to *A. niger* compared to the two Basidiomycetes, and has a much better orthology-based functional prediction. Based on this, *A. niger* cannot fully serve as a reference for other Basidiomycetes. It would therefore be needed to generate a separate reference model for Basidiomycetes, and the availability of high-quality reference genomes for *P. chrysosporium* (https://mycocosm.jgi.doe.gov/Phchr4_2/Phchr4_2.home.html) and/or *D. squalens* (https://mycocosm.jgi.doe.gov/Dicsqu464_2/Dicsqu464_2.home.html) could be used to achieve this, significantly expanding the toolbox for efficient metabolic engineering of these species. These resources contribute to our understanding of lignocellulose conversion mechanisms, enable comparative genomics analyses, and provide avenues for the development of more sustainable and efficient biotechnological applications.

In addition, in **Chapter 3**, I employ transcriptome, proteome and metabolome

data to comprehensively analyze the response of sugar metabolic genes to different monosaccharide conditions. Comprehensive analysis of three omics data shows that sugar metabolic genes of closely related species (four Ascomycetes) show similar expression or abundance patterns on the corresponding sugar conditions. Similar analyses have been done for individual species [2-8] and comparisons have been made in their discussions. However, this was the first time a comparison of highly diverse species was performed using an identical approach and a direct gene comparison based on orthology, providing a much deeper understanding of the species-specific approaches, with respect to genome content and gene expression.

### *1.1 Comparing our models with the KEGG database*

KEGG (Kyoto Encyclopedia of Genes and Genomes) PATHWAY [9], established in 1995, contains maps for various metabolic pathways for a variety of organisms. It is used for a variety of applications including genome analysis and metabolic engineering. As described in Table S1.1, only 120 fungal species are available in the KEGG database, including 97 Ascomycetes and 22 Basidiomycetes. With regards to carbohydrate metabolism, the KEGG database divides carbohydrate metabolism into many categories (Table S1.2), of which several important metabolic pathways (highlighted in Table S1.2) are not well classified for fungi.

The studies described in this thesis went beyond the KEGG database with respect to accuracy and completeness of fungal primary metabolism. Our comparison between the sugar metabolism models of *A. niger*, *A. nidulans*, *T. reesei*, and *D. squalens* and the information available in the KEGG database demonstrated that our models offer greater specificity and depth in representing sugar metabolism of these fungi.

Firstly, we compiled a comprehensive list of associated reactions for each monosaccharide. However, we found that the KEGG database does not provide this specific information. For instance, within the Pentose Phosphate

Pathway (PPP), we observed several reactions related to glycolysis or the gluconate pathway (Table S1.4). These include glycolysis-related EC numbers (4.1.2.13, 3.1.3.11, and 2.7.1.11) and gluconate pathway-related EC numbers (3.1.1.17 and 2.7.1.12). In addition, we observed the absence of crucial reactions related to L-arabinose metabolism (such as EC numbers 1.1.1.21, 1.1.1.12, and 1.1.1.10) in the KEGG database (Table S1.5), as well as some reactions within the D-galactose pathways (Table S1.7). Furthermore, it was noted that the KEGG database includes certain unrelated reactions within the glycolysis pathway (as shown in Table S1.3). Similar occurrences were observed in other pathways as well, as outlined in Table S1.

Secondly, the KEGG database does not consistently update the genome annotation information for each species and may still include pseudogenes that have been eliminated in the latest genome version. This could potentially mislead researchers and provide inaccurate guidance during searches. The most obvious example is that in *A. nidulans* there are many genes that no longer exist in the genome version currently available, e.g. AN1735 involved in PPP (Table S1.4), as well as AN4828 and AN3516 in D-galactose metabolism (Table S1.7). Besides, the annotations of many genes in *A. nidulans* were derived from the deprecated Aspergillus Genome Database (AspGD) database [10], such as the genes involved in starch and sucrose metabolism in KEGG database, but their annotations are related to maltose metabolism in the AspGD database.

Last but not least, no specific pathways were assigned for fungal L-rhamnose and D-galacturonate metabolism in the KEGG database. Instead, these two metabolic pathways are subsumed into fructose and mannose metabolism, and pentose and glucuronate interconversions, respectively. So, we propose that for fungi KEGG can be a starting point to evaluate metabolism, but it clearly does not represent the full scale of pathways and related genes/enzymes that is currently known in fungi. As an increasing number of fungal genomic studies uses KEGG as one of their annotation tools [8, 11-17], more

awareness of the limitations of KEGG and its accuracy should be raised in the fungal community. Limited publications so far address this [18] and suggest improvements, such as those of the bioreaction database [19, 20], which is unfortunately not available online.

### *1.2 Considerations for functional separation*

In **Chapter 3** of this thesis, we described the extrapolation of sugar metabolism of *A. niger* to different species by using an orthology-based approach. In fact, the increasing taxonomic distance of these species to *A. niger* likely resulted in an increasing error rate with respect to orthology-based functional predictions. In certain instances, we may encounter a scenario where multiple orthologs correspond to a single sugar metabolic gene, making it challenging to readily identify or select authentic and reliable orthologs in other fungi. To solve this, phylogeny (based on amino acid sequence similarity) is an effective method that can help us with functional classification and true ortholog identification. A phylogenetic tree, commonly referred to as a phylogeny, is a graphical representation illustrating the evolutionary relationships and ancestral lineage of various species, organisms, or genes derived from a shared ancestor [21]. Basically, phylogenetic tree generation consists of sequence alignment where the resulting tree reveals how alignment can influence the tree formation. Alignment-based methodologies are probably the most widely used tools in sequence analysis problems. In our experiments, Multiple Alignment using Fast Fourier Transform (MAFFT, https://mafft.cbrc.jp/alignment/server/) is the most reliable tool for the alignment. Sequences obtained from different fungal genomes (e.g., from the JGI MycoCosm database, https://mycocosm.jgi.doe.gov/mycocosm/home) are combined with the characterized genes (such as sugar metabolic genes in *A. niger*) to build phylogenetic trees. In the phylogenetic tree, the classification status of genes without functional description can be clearly inferred based on the functions of known genes. However, when applying this method for functional separation, we must take into account an additional factor - the high error rate of fungal gene models.

Incorrect gene models are commonly present in fungal genomes even if fungal genomes with good quality are selected. These errors primarily occur in the prediction of the start codon (ATG), stop codon (TAG, TAA, and TGA), and in the intron-exon boundaries [22]. Thus, reliable phylogenies should be based on correct amino acid sequences, and manual curation of gene models, even though this is time consuming, is crucial to achieve a correct annotation and reliable phylogenetic analyses.

## 2. The diversity of CAZy families involved in plant biomass degradation

### *2.1 Genomic diversity in plant biomass degradation*

Before the genome era, many of the in-depth studies focused on a relatively small number of species, whereas now, many species can be addressed in detail, revealing the large variety in the approach used by fungi to degrade plant biomass. Diversity of CAZy content in fungal genomes has been observed in many studies [23-26], although not all of them compared this down to the individual gene function or orthology. In **Chapter 4**, I compared five fungi with increasing taxonomic distance to *A. niger* in detail, revealing the large variety in the approach to degrade plant biomass. We observed variation at many levels, including genomic adaptation to the preferred biomass component as well as distinct variation in sets of CAZy families encoded in their genomes related to individual polysaccharides. Specifically, among the studied fungi, the Eurotiomycetes exhibit a higher abundance of CAZy genes associated with pectin degradation, whereas the two basidiomycetes possessed a larger number of CAZy genes involved in cellulose degradation (**Chapter 4**). However, we should not ignore the large CAZy content variability within Basidiomycete genomes, especially when comparing species with either white-rot, brown-rot, or grey-rot life styles [27]. *T. reesei* is a special case because it contains a reduced enzyme set but it has been shown that it produces these at high levels, especially cellulolytic enzymes [28, 29]. Other *Trichoderma* species do not share this contraction, but have a more comparative CAZy

content to other Sordariomycetes, based on a recent genomic comparison of >20 *Trichoderma* species (unpublished data). Additionally, larger differences were detected through transcriptome analysis, even between closely related species, suggesting a high level of adaptation in individual species. As summarized in **Chapter 4**, among the Eurotiomycetes, distinct sugar-specific expression patterns were observed. For instance, *A. niger* and *P. subrubescens* have a larger number of CAZy genes with high sugar specific expression on L-arabinose. In contrast, *A. nidulans* demonstrated an abundance of CAZy genes primarily associated with L-rhamnose induction. Previously, high variability between species of the same genus was also observed for *Aspergillus*, *Penicillium*, *Meliniomyces*, *Phanerochaete* and *Talaromyces* [22], demonstrating the high evolutionary changes in this biological process. The two Basidiomycetes differed significantly from the Ascomycetes in that they displayed moderate sugar-specific expression patterns. Possibly this is due to the lack of orthologs of the Ascomycetes transcriptional activators related to this process in Basidiomycetes [30], which suggests a differently organized activation of these CAZy genes between the two phyla. It would also be interesting for future studies to study the diversity of CAZy content and the presence of such regulators in other fungal phyla, such as the Mucoromycetes, which are also highly relevant for biotechnology.

### *2.2 The value of whole genus genome sequencing projects*

Recent insights from my studies also revealed the high value that can be obtained from whole genus genome sequencing projects that are summarized here. The genome mining strategy has become increasingly popular with the explosion of available fungal genome sequences. In recent years, new initiatives have emerged that address genome sequencing of all species of fungal genera. Currently, three of such projects are running through JGI addressing the biotechnologically relevant genera *Aspergillus* [31-34], *Penicillium* and *Trichoderma* (https://jgi.doe.gov/csp-2018-berka-genus-wide-genomics-trichoderma/).

Fungal genomics has revolutionized our understanding of the diversity of fungi with respect to their enzymatic potential, and has unveiled substantial variations in the number of genes for specific CAZy families [22]. As mentioned in **Chapter 4**, we already know that even closely related fungi deploy diverse enzymatic strategies to degrade plant biomass. Thus, we would like to explore the diversity of CAZy families in larger sets of closely related fungal genomes, such as across the *Aspergillus* genus. The *Aspergillus* project is close to completion, which provides us with a good starting point, so we evaluated a selection of 196 of these genomes for CAZy diversity and evolution, which are classified into 24 sections, and also added four *Penicillium* genomes as outgroup (Table S2). Despite belonging to the same *Aspergillus* genus, the CAZy gene content among these 196 fungi exhibits significant variation, ranging from less than 50 to over 300 genes. Even within the same taxonomic sections (a taxonomic rank below the genus), we observed that the distribution of CAZy content across species within the same section is not concentrated as expected, especially in Fumigati, Nidulantes and Nigri (Figure 1).

However, when we go through the CAZy genes content involved in each plant biomass substrate, we found that fungi within the same section had overall similar proportions of CAZymes involved in different substrates although their total enzymes were variable (Figure 2). Interestingly, within certain sections such as Nidulantes, Nigri, and Fumigati (and several others), significant variations in scale are observed, indicating notable differences in CAZy content among these species. In a recent analysis of another fungal genus, *Trichoderma*, much lower variation of CAZy content in its sections was observed (unpublished data).

In addition, as expected, it was observed that the variation between strains is extremely minimal, such as the six *A. niger* strains included in this study (Figure 2). For two of these *A. niger* strains included in this study, it was previously shown that one abundantly produced cellulolytic enzymes, while

the other produced hardly any [35, 36]. While the evolutionary drivers for this diversity remain to be elucidated, it has been shown that expansion or reduction of enzymes related to a certain polysaccharide correlates with improved or reduced growth of the fungus on that polysaccharide [22, 34, 37].

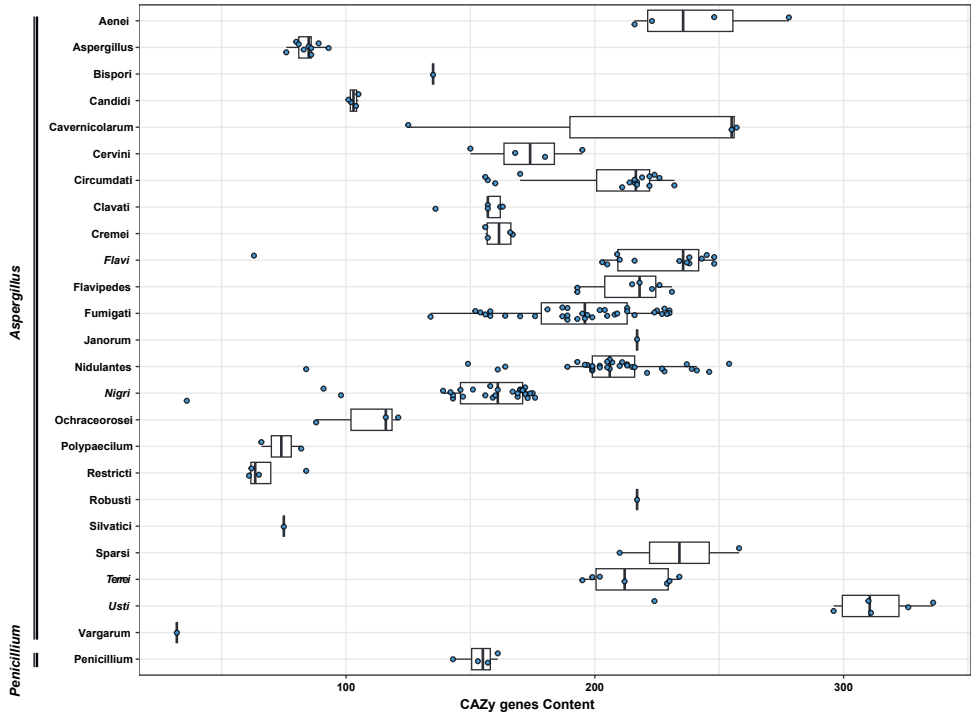


**Figure 1. Distribution of CAZy gene content among different species within each section of the genus *Aspergillus***. Each blue dot represents a fungal species or strain. The box displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum, reflecting the distributions of the number of CAZy genes within each section.

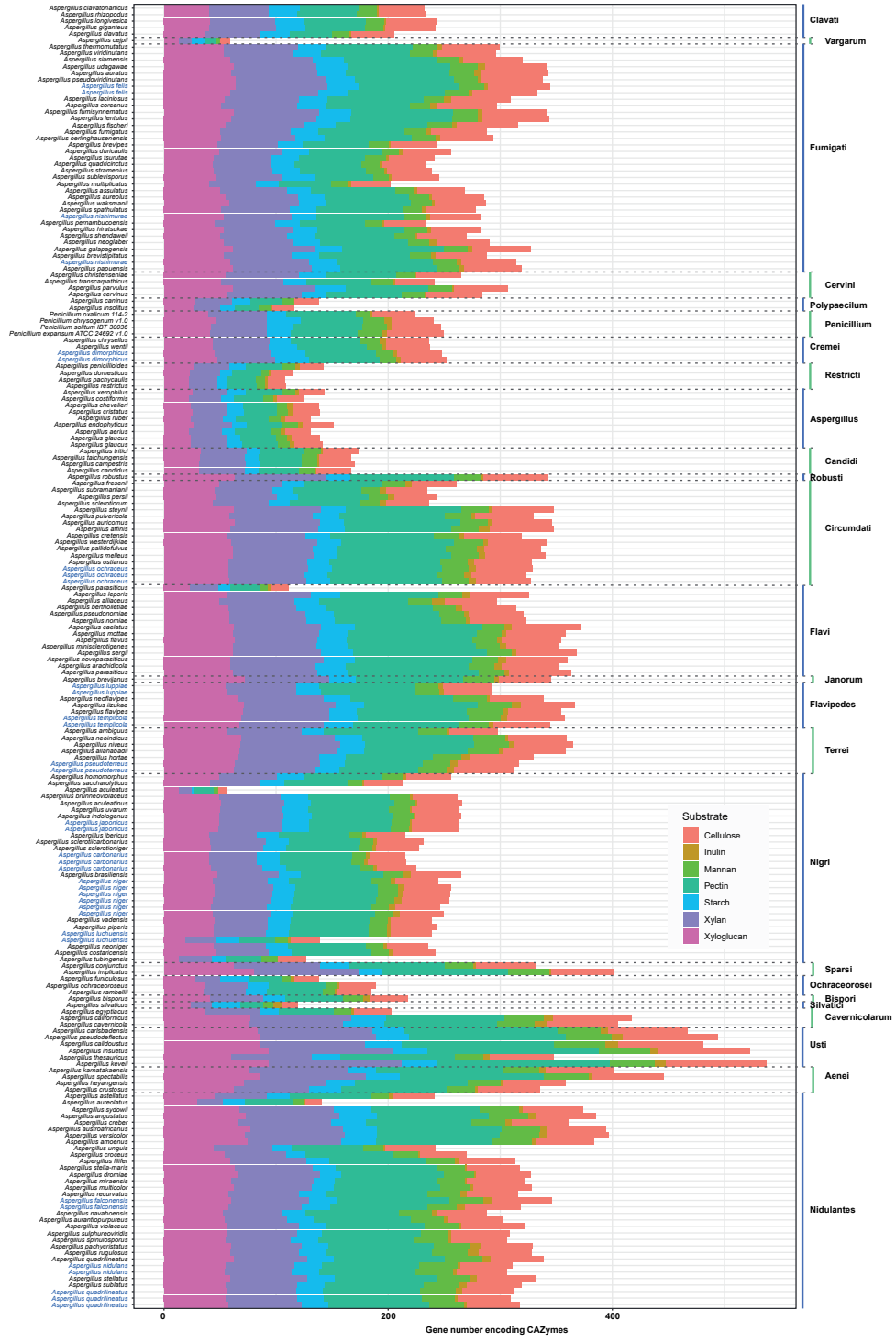
Substrate
Cellulose
Inulin
Mannan
Pectin
Starch
Xylan
Xyloglucan
0
200
400
Gene number encoding CAZymes
Clavati
Vargarum
Fumigati
Cervini
Polypaecilum
Penicillium
Cremei
Restricti
Aspergillus
Candidi
Robusti
Circumdati
Flavi
Janorum
Flavipedes
Terrei
Nigri
Sparsi
Ochraceorosei
Bispori
Silvatici
Cavernicolarum
Usti
Aenei
Nidulantes

**Figure 2. Summary of the number of CAZy genes involved in individual plant polysaccharide in the 200 studied fungi**. The color of the bar indicates different substrates. The name of fungi highlighted in blue indicates multiple strains of the corresponding species. The color of the bar indicates different substrates.

As mentioned previously, Aspergilli have extensive sets of CAZy genes that are crucial for the degradation of plant biomass. Therefore, it is essential to examine the diversity within the genus at both the genotype and phenotype levels to gain a comprehensive understanding of this trait. Based on the growth profiles of 78 *Aspergillus* species on 36 plant biomass-related carbon sources, I observed that species within the same section that possess similar numbers of putative enzymes, have significantly different growth profiles related to the corresponding CAZyme content (Table S3). For example, five species from Circumdati section displayed quite different growth profiles. Although the enzyme content involved in xylan degradation was comparable, significant variations were noted in the growth profiles of the two types of crude xylan substrates. Interestingly, these differences correlated with the growth profiles observed on L-arabinose and D-xylose, further highlighting the impact of specific sugar metabolic pathways on fungal growth. The similar scales are seen in the Nidulantes and Nigri sections. Through such a pilot project, we learned that while actual carbon utilization is complex, and understanding these mechanisms solely based on gene content is insufficient, it does provide us with some initial clues and a foundation for further exploration.

This post-genomic diversity was also observed in to previous studies [35, 36], in which the proteomic response of a number of Aspergilli was compared during growth on plant biomass. Despite similar CAZy genome content, the enzymes that were produced differed significantly, even between the two *A. niger* strains that were included in these studies. This also indicates that comparing plant biomass degrading ability of fungi, purely based on their genome content, may not reflect their true abilities, as the regulation of the expression of the corresponding genes appears to affect the produced enzyme set even more.

As mentioned earlier, more whole genus genome projects will follow. Whole genus genome projects will help elucidate the genomic diversity and evolutionary dynamics within a genus, like we did in the pilot project. By analyzing variations in gene content, gene family expansions or contractions, and genetic rearrangements, we could infer the mechanisms driving genome evolution and adaptation within the genus. In addition, it will contribute to taxonomic classification, phylogenetic analysis, and comparative genomics studies. For example, comparative genomics can reveal conserved genes, genomic features, and regulatory elements, as well as identify species-specific or genus-specific genes and genomic variations. These comparative analyses provide valuable insights into the genetic foundations underlying unique traits, adaptations, and evolutionary processes exhibited by different species. The low taxonomic distance and overall variability in genome content makes such genome sets ideally suited for machine learning approaches to predict growth phenotypes of a wide range of species based on a characterized reference set. Finally, these projects will also allow comparison of the extent of variation within different general as well as the correlation of genome content to lifestyle (e.g. saprobe, symbiont, pathogen) for those species for which this is known and possible prediction of life style for those for which this is not yet known. Overall, whole genus genome sequencing projects provide a wealth of data that contribute to our understanding of genomics, evolution, taxonomy, and biology within a particular genus.

## 3. Insights into fungal postgenomic research

The development of postgenomic, e.g. transcriptomic, proteomic, and metabolomic allows us to better track the diversity of fungal plant biomass degradation approaches, from the transcript level, over the protein level, to the metabolite level. In **Chapter 2**, I systematically introduced the principles of different omics data and their applications in fungal research. Although many transcriptome, proteome, and metabolome studies during the growth of fungi on plant biomass have been reported [38], there still are certain

considerations. Transcriptomic analysis captures gene expression at a specific moment, providing a snapshot of the transcript levels. It allows us to capture and analyze gene expression changes quickly, observe immediate responses to different experimental conditions, and gain insights into transient or dynamic biological processes. In these cases, we need to design transcriptomic experiments to capture specific temporal events or transitions in gene expression. For example, time-course experiments are designed to track gene expression changes over a defined period, allowing us to study dynamic processes such as response to a specific treatment [39]. For fungal research, many expression profiles of a "large" time course of CAZy and pathway genes have been reported in Aspergilli [40-42], *T. reesei* [43], *Neurospora crassa* [44], and *Myceliophthora thermophila* [45] or mixed cultures [30, 46]. However, transcript abundance does not reveal the substantial role of the regulatory mechanisms occurring after transcription as protein abundance only partially can be explained by transcript abundance [47, 48].

In contrast, proteomics provides a more comprehensive view of protein production and activity compared to transcriptomics. Proteomic analysis could cover the methods to determine the identification and quantitation of proteins as well as protein-protein interactions, protein modifications and localization within cell or tissue. For example, posttranslational modifications may influence protein activity and function, and more studies on this field were reported [49-51]. In this respect mass spectrometry techniques are very useful as they can characterize and quantify posttranslational modifications. However, this process necessitates the proper separation and preparation of biological samples to effectively detect modifications of interest [47]. Proteomics techniques also do not provide complete coverage of the proteome so that certain proteins may be challenging to detect or quantify [52, 53], especially low-abundance proteins. Thus, the detection and quantification of both highly abundant and low-abundance proteins in the same sample pose technical challenges. In addition, accurate identification and annotation

of proteins from mass spectrometry data can be challenging. Appropriate bioinformatics tools and accurate protein sequence databases both are essential for protein identification [54]. For example, incomplete or inaccurate databases, particularly for non-model organisms and mis-annotated gene models, can lead to misidentification or non-identification of proteins.

Similarly, metabolomics techniques also cannot detect the entire metabolome [55] because some metabolites may be present at low concentrations, or remain undetected, or not fully characterized, or have properties that make them difficult to detect using current analytical methods. It should also be recognized that only metabolites that are part of the reference database for the mass spectrometer used will be identified, while all others will be labelled as unknown. In addition, the stability of metabolites could be another complication. This can vary depending on several factors, such as sample collection and storage conditions, and the analytical techniques used. Therefore, it is important to carefully optimize sample handling and storage protocols to minimize changes in metabolite composition and ensure data stability. A fundamental advantage of metabolomics data is that it can be reused for various purposes and analyses [56, 57]. After the data is acquired, it can be analyzed using different bioinformatics tools to derive additional insights and explore different research questions or integrate with other omics datasets. In terms of databases, many specific databases are available for metabolomics, which could provide a reference for metabolite identification, pathway analysis, and data integration, facilitating the interpretation and reuse of metabolomics data. However, these databases can often not be transferred between different types of mass spectrometry instruments. While databases provide valuable information, there are still many unknown or poorly characterized metabolites, which undoubtedly brings challenges to identify and annotate metabolites. Therefore, efforts to expand metabolomics databases and improve metabolite annotation methods will contribute to better metabolite identification and characterization.

In summary, to gain a more comprehensive understanding of biological mechanisms, integrating multiple omics approaches such as transcriptomics, proteomics, and metabolomics can provide a more holistic view of biological processes.

## 4. Phenotypic differences between the species

Although we observed significant diversities across different fungal species in genomic and post-genomic levels, these differences did not always have a clear correlation with growth profiles on plant biomass-related carbon sources. In **Chapter 3**, the six fungi showed notable similarities and differences with respect to growth on nine basic monosaccharides involved in sugar metabolism. Overall, it has been shown that the absence or presence of sugar metabolic genes related to a certain sugar metabolic pathway directly affects fungal growth on the corresponding monosaccharides. An obvious example illustrating this phenomenon is the contrasting growth on L-arabinose observed between Ascomycetes and Basidiomycetes. The limited growth of L-arabinose observed in two Basidiomycetes species can potentially be attributed to the absence of key genes involved in the pentose catabolic pathway (**Chapter 3**). However, this is not the case for the growth on D-galactose. Despite the presence of the necessary genes for the D-galactose pathway in the genomes of *A. niger* and *T. reesei*, both fungi exhibit remarkably poor growth on D-galactose. For *A. niger*, it was shown that the inability to grow on D-galactose as the sole carbon source from spores was linked to an absence of D-galactose import by conidiospores, while already growing mycelium was able to import D-galactose [58]. Interestingly, this seems to be a common feature among fungi as all black Aspergilli, except *A. brasiliensis*, share this phenotype [59] as do many other fungal species. In **Chapter 4**, growth of the six species (described in **Chapter 3**) was evaluated on 18 plant biomass related carbon sources. The data suggest that the variation in growth between taxonomic distant species could be linked to the differences in CAZy gene content, but not always. Similar patterns were also found in previous

comparative genomic studies [37, 60-63]. An illustrative example is the observation that the reduction of CAZyme content in the genome of *T. reesei* roughly correlated with its poor growth on most of the tested polysaccharides. Similarly, the absences of the crucial CAZy genes of GH32 in *A. nidulans*, *T. reesei* and two Basidiomycetes result in the diminished growth on inulin. However, this was not the case for closely related species, as highlighted in our aforementioned pilot project, thus reinforcing previous findings [31-33]. In summary, the genomic potential and variations are not necessarily reflected in the growth. It is therefore likely that as suggested previously, the observed variation is mainly caused by regulatory changes, rather than simply by gene content [34, 64].

## 5. Exploration of co-evolution across the fungal kingdom

Fungi have evolved a sophisticated system to efficiently degrade and utilize plant biomass (as depicted in Figure 1 in **Chapter 1**). In **Chapter 5**, a comparative transcriptome analysis was conducted on five different fungi (*A. niger*, *A. nidulans*, *P. subrubescens*, *T. reesei*, and *P. chrysosporium*), focusing on the transcriptome response of CAZymes, sugar transporters, and sugar metabolic genes on two crude substrates (corn stover (CS) and soybean hulls (SBH)) at three distinct time points. The results of the transcriptome analysis revealed distinct time, substrate, and species-specific patterns in the fungal adaptation to plant biomass, highlighting diverse approaches of these fungal species for adaptation to crude plant materials and their complex transcription regulation. Previous studies on individual fungi also demonstrated such patterns [5, 43, 65-67], but this was the first time these were directly compared between distant fungal species. In addition, an interesting observation was made during this study, as CAZy genes, sugar metabolic genes, and sugar transporters exhibited similar or consistent expression trends. This finding sparked my thinking in exploring the potential co-evolutionary patterns related to plant biomass degradation.

In theory, if gene expansion/reduction (or absence/presence) of CAZy genes,

sugar metabolic genes and sugar transporters occurs at the same time, it can be based on this genetic information to explain that they have co-evolved. This would be expected as evolving random gene content would not be sufficient to modify the approach of such a complex biological system. Here, I selected 19 fungi from different phyla to compare their potential evolutionary trends (Figure 3 and Table S4). As far as one of the three gene categories is concerned, they generally show a trend consistent with the their evolutionary, i.e., more distant species have smaller gene repertoires. This is consistent with our previous results (**Chapter 3** and **Chapter 4**). However, we also noticed some features of phylum-specificity or species-specificity. In terms of species specificity, CAZy genes involved in xyloglucan degradation were almost absent in *Neurospora crassa*, and the GH32 family was absent in *Agaricus bisporus*. Surprisingly, we did observe growth for *A. bisporus* on inulin (Table S5), suggesting that it has employed other enzymes for its degradation. In terms of phylum-specificity, only six Ascomycetes of this set of 19 fungi possess genes related to sucrose metabolism (Figure 3). When considering the integrated comparison of these three gene types, we observe a general pattern of co-evolution. For example, Basidiomycetes have fewer CAZy genes and sugar transporters.

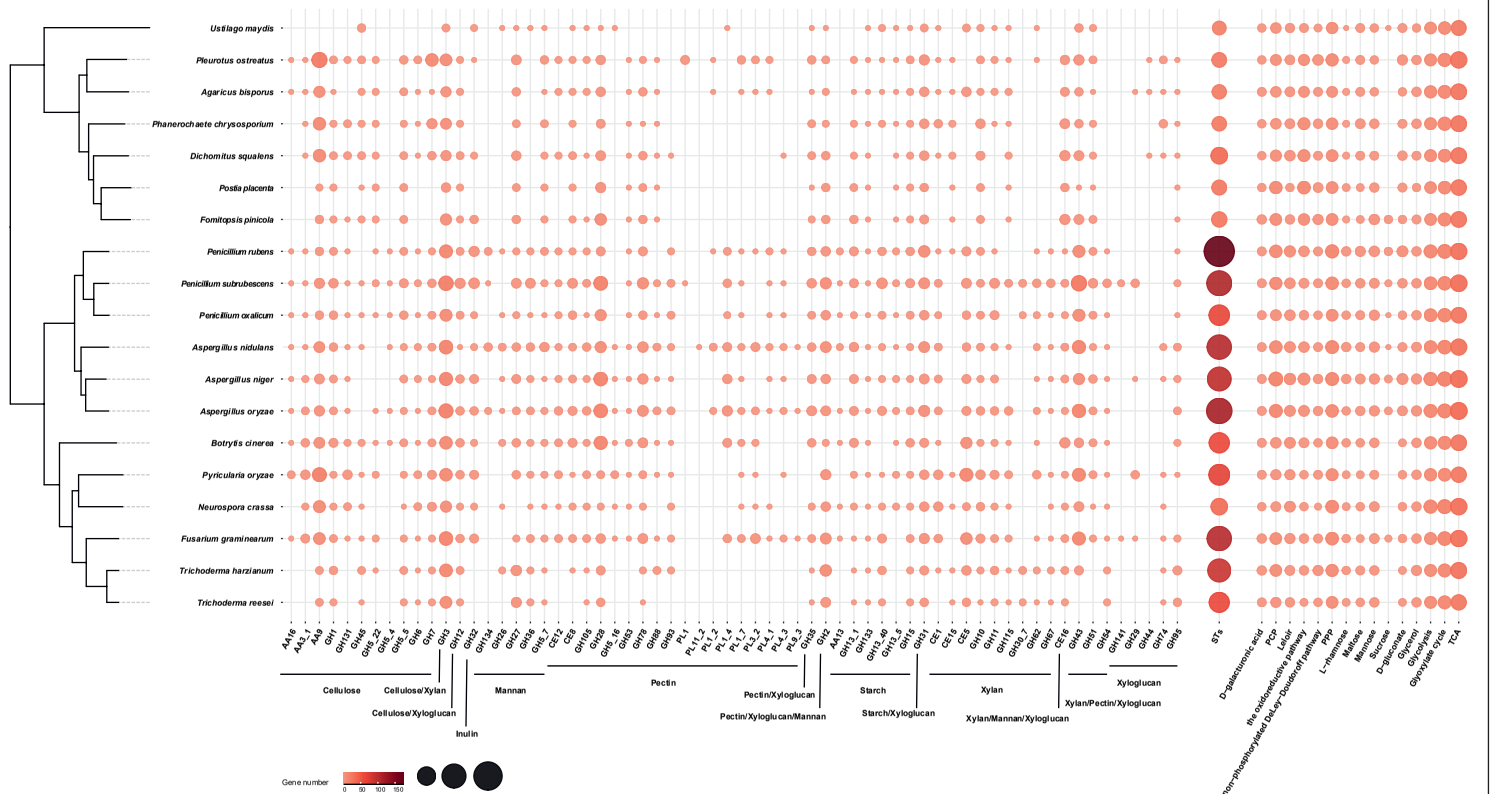


**Figure 3. Distribution of CAZymes, sugar transporters and sugar metabolic genes in 19 fungi**. CAZymes acting on the same polysaccharide were grouped together and were highlighted with individual lines. The size and color of the dots indicates the number of the corresponding genes.

## 6. Concluding remarks

Taken together, in this thesis, the importance of gaining a deeper understanding of the molecular mechanisms underlying fungal plant biomass conversion is emphasized. The thesis starts by exploring the application of bioinformatics in fungal research, providing an overview of the bioinformatics methods and software used in the field (**Chapter 2**). These methods are continuously evolving but still require further improvement. One example of this are the gene model prediction algorithms. The high percentage of erroneous gene models in fungal genomes have a major impact on research in this field [68]. They not only affect phylogenies and protein production by gene synthesis, but also provide an incorrect reference for transcriptomics and proteomics, causing an incorrect quantification of the levels of gene expression or protein production. Better algorithms for gene model prediction are therefore urgently needed to reduce these problems.

**Chapter 3** delved into the possibility of transferring the fungal central carbon metabolism from *A. niger* [69] to other fungi. Such models significantly contribute to the understanding of the conservation of the catabolic pathways and the order in which sugars are converted across the fungal kingdom. My study demonstrated that the transfer of this model is only reliable within a fungal phylum. Therefore, I suggest that a *de novo* model should be created for each major fungal phylum and then be tested on several other species of that phylum to evaluate its transferability. It would also be highly interesting and relevant to expand the current model to involve not only primary sugar metabolism, but link it to other important parts of metabolism, such as amino acid metabolism and secondary metabolism. This would reveal to which extent the carbon source affects those metabolic pathways, moving closer to a full systems biology understanding of fungal physiology.

**Chapter 4** presents a comparative analysis of genome content and expression induction patterns across six fungi. This analysis provides a comprehensive understanding of the diversity of plant biomass degradation at the genomic

and transcriptomic levels. By comparing the genomes and studying gene expression patterns, we can unravel the genetic basis for the variations observed in plant biomass degradation among different fungal species. The strong differences between them confirm previous studies done for a set of Aspergilli [36] and highlight that similar genome content does not imply a similar approach. The approach seems to be more dependent on the regulation of the relevant genes and considering the only partial conservation of transcriptional regulators related to plant biomass conversion in fungi [30], this creates already significant differences between species. In addition, it has been shown that the same regulator only has a similar, but not an identical role in different species [64, 70, 71]. Care should therefore be taken when comparing the ability of fungi for plant biomass conversion solely on genome content. Similarly, the presence of an ortholog of a known regulator does not automatically mean that the regulation of genes can be predicted.

This was further exemplified in **Chapter 5** in which the transcriptome profile of genes encoding CAZymes, sugar transporters, and sugar metabolic enzymes was analyzed on two crude substrates and combined with the previously established transcriptome response on nine monosaccharides. This analysis highlights that different fungal species employ diverse strategies for adapting to and utilizing specific plant materials, and reveals species-specific patterns, indicating variations in the expression of CAZymes, sugar transporters, and sugar metabolic genes among the different fungi. Plant biomass conversion is a crucial process for many fungi in their natural environment as this substrate is a major carbon and energy source. It is therefore not surprising that a sophisticated regulatory system controls the expression of diverse gene sets that have evolved to match very well to the substrates a fungus typically encounters in its biotope. A better understanding of these different approaches has enormous potential for the improvement of existing and design of new applications for fungi in the development of the biobased economy.

The overall objective of this thesis was to contribute to the advancement of

knowledge in the field of fungal plant biomass conversion. By integrating bioinformatics approaches, examining sugar metabolism and CAZymes, and conducting comparative genomic and transcriptomic analyses, this research aimed to provide valuable insights into the mechanisms and diversity of fungal plant biomass degradation. The two main overall results of my project in this context relate to the comparison of the plant biomass conversion abilities of fungi based on genome information alone. The study described in **Chapter 3** demonstrates clear limitations to homology-based assumptions across the fungal tree of life. Within relatively small taxonomic distances (within a phylum) homology-based functional assignment of individual genes as well as pathways/systems is quite reliable, but this reliability is reduced strongly when applying this approach beyond the phylum boundaries. I would therefore recommend the establishment of a reference species/genome per phylum that has good experimental support.

Secondly, **Chapter 4** demonstrates that comparing plant biomass degrading ability of fungi on genome content alone does not provide an accurate situation. Conserved genes between species can have significantly different expression patterns, indicating the regulation of gene expression may more strongly affect the plant biomass degrading ability than the genome content itself. This can be partially explained by the presence/absence of relevant transcriptional activators in fungal species, but is also due to the difference in functioning and target gene set of conserved transcriptional activators. Genome-content based comparisons therefore should be interpreted more cautiously than has often been the case. Finally, **Chapter 5** demonstrates the strong differences when comparing the expression of genes involved in plant biomass conversion on crude plant biomass to pure monosaccharides. The complexity of natural substrates results in complex expression patterns that are not just combinations of the monosaccharide profiles, and that show different relationships between them in different species. Additional research should be dedicated to better understanding the complex molecular mechanism underlying the expression

profiles on crude substrates, as these are the natural substrates of fungi in their biotopes and also often the ones used in industrial applications. A multi-omics approach would be highly beneficial for this, as it allows an evaluation of gene expression together with steady-state levels of enzymes and metabolites, resulting in a more system-biology based understanding of the plant biomass conversion process.

## References


1. Li, J., et al., *The sugar metabolic model of Aspergillus niger can only be reliably transferred to fungi of its phylum.* **Journal of Fungi**, 2022. 8(12): p. 1315.
2. Zhang, C., et al., *First report on the regulation and function of carbon metabolism during large sclerotia formation in medicinal fungus Wolfiporia cocos.* **Fungal Genetics and Biology**, 2023. 166: p. 103793.
3. Li, J., et al., *Transcriptional comparison of the filamentous fungus Neurospora crassa growing on three major monosaccharides D-glucose, D-xylose and L-arabinose.* **Biotechnology for Biofuels**, 2014. 7(1): p. 1-15.
4. Peng, M., et al., *CreA-mediated repression of gene expression occurs at low monosaccharide levels during fungal plant biomass conversion in a time and substrate dependent manner.* **The Cell Surface**, 2021. 7: p. 100050.
5. Sun, J., et al., *Deciphering transcriptional regulatory mechanisms associated with hemicellulose degradation in Neurospora crassa.* **Eukaryotic Cell**, 2012. 11(4): p. 482-493.
6. David, H., et al., *Metabolic network driven analysis of genome-wide transcription data from Aspergillus nidulans.* **Genome Biology**, 2006. 7: p. 1-16.
7. Patyshakuliyeva, A., et al., *Carbohydrate utilization and metabolism is highly differentiated in Agaricus bisporus.* **BMC Genomics**, 2013. 14: p. 1-14.
8. Li, T., et al., *Comparative transcriptome analysis of Penicillium citrinum cultured with different carbon sources identifies genes involved in citrinin biosynthesis.* **Toxins**, 2017. 9(2): p. 69.
9. Kanehisa, M. and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes.* **Nucleic Acids Research**, 2000. 28(1): p. 27-30.
10. Arnaud, M.B., et al., *The Aspergillus Genome Database (AspGD): recent developments in comprehensive multispecies curation, comparative genomics and community resources.* **Nucleic Acids Research**, 2012. 40(D1): p. D653-D659.
11. Hirpara, D.G., et al., *Exploring conserved and novel MicroRNA-like small RNAs from stress tolerant Trichoderma fusants and parental strains during interaction with fungal phytopathogen Sclerotium rolfsii Sacc.* **Pesticide Biochemistry and Physiology**, 2023. 191: p. 105368.

12. Wu, T., et al., *Whole-genome sequencing and transcriptome analysis of Ganoderma lucidum strain Yw-1-5 provides new insights into the enhanced effect of Tween80 on exopolysaccharide production.* **Journal of Fungi**, 2022. 8(10): p. 1081.
13. Liu, N., et al., *Integrative transcriptomic-proteomic analysis revealed the flavor formation mechanism and antioxidant activity in rice-acid inoculated with Lactobacillus paracasei and Kluyveromyces marxianus.* **Journal of Proteomics**, 2021. 238: p. 104158.
14. Pan, L., et al., *Dimethylformamide inhibits fungal growth and aflatoxin b1 biosynthesis in Aspergillus flavus by down-regulating glucose metabolism and amino acid biosynthesis.* **Toxins**, 2020. 12(11): p. 683.
15. Zhao, X., et al., *GC–MS-based nontargeted and targeted metabolic profiling identifies changes in the Lentinula edodes mycelial metabolome under high-temperature stress.* **International Journal of Molecular Sciences**, 2019. 20(9): p. 2330.
16. Xiao, C., et al., *Hypoglycemic mechanisms of Ganoderma lucidum polysaccharides F31 in db/db mice via RNA-seq and iTRAQ.* **Food & Function**, 2018. 9(12): p. 6495-6507.
17. Wang, J., et al., *Analysis of ethanol fermentation mechanism of ethanol producing white-rot fungus Phlebia sp. MG-60 by RNA-seq.* **BMC Genomics**, 2016. 17(1): p. 1-11.
18. Altman, T., et al., *A systematic comparison of the MetaCyc and KEGG pathway databases.* **BMC Bioinformatics**, 2013. 14(1): p. 1-15.
19. Ma, H. and A.-P. Zeng, *Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms.* **Bioinformatics**, 2003. 19(2): p. 270-277.
20. Stelzer, M., et al., *An extended bioreaction database that significantly improves reconstruction and analysis of genome-scale metabolic networks.* **Integrative Biology**, 2011. 3(11): p. 1071-1086.
21. Jill Harrison, C. and J.A. Langdale, *A step by step guide to phylogeny reconstruction.* **The Plant Journal**, 2006. 45(4): p. 561-572.
22. de Vries, R.P. and M.R. Mäkelä, *Genomic and postgenomic diversity of fungal plant biomass degradation approaches.* **Trends in Microbiology**, 2020. 28(6): p. 487-499.
23. Barrett, K., et al., *Fungal secretome profile categorization of CAZymes by function and family corresponds to fungal phylogeny and taxonomy: Example Aspergillus and Penicillium.* **Scientific Reports**, 2020. 10(1): p. 5158.
24. Challacombe, J.F., et al., *Genomes and secretomes of Ascomycota fungi reveal diverse functions in plant biomass decomposition and pathogenesis.* **BMC Genomics**, 2019. 20(1): p. 1-27.

25. Pellegrin, C., et al., *Comparative analysis of secretomes from ectomycorrhizal fungi with an emphasis on small-secreted proteins.* **Frontiers in Microbiology**, 2015. 6: p. 1278.
26. Park, Y.-J., Y.-U. Jeong, and W.-S. Kong, *Genome sequencing and carbohydrate-active enzyme (CAZyme) repertoire of the white rot fungus Flammulina elastica.* **International Journal of Molecular Sciences**, 2018. 19(8): p. 2379.
27. Rytioja, J., et al., *Plant polysaccharide degrading enzymes from basidiomycetes.* **Microbiology and Molecular Biology Reviews**, 2014. 78(4): p. 614-649.
28. Peterson, R. and H. Nevalainen, *Trichoderma reesei RUT-C30–thirty years of strain improvement.* **Microbiology**, 2012. 158(1): p. 58-68.
29. Persson, I., F. Tjerneld, and B. Hahn-Hägerdal, *Fungal cellulolytic enzyme production: a review.* **Process Biochemistry**, 1991. 26(2): p. 65-74.
30. Benocci, T., et al., *Regulators of plant biomass degradation in ascomycetous fungi.* **Biotechnology for Biofuels**, 2017. 10(1): p. 1-25.
31. Kjærbølling, I., et al., *A comparative genomics study of 23 Aspergillus species from section Flavi.* **Nature Communications**, 2020. 11(1): p. 1-12.
32. Vesth, T.C., et al., *Investigation of inter-and intraspecies variation through genome sequencing of Aspergillus section Nigri.* **Nature Genetics**, 2018. 50(12): p. 1688-1695.
33. Benoit, I., et al., *Closely related fungi employ diverse enzymatic strategies to degrade plant biomass.* **Biotechnology for Biofuels**, 2015. 8(1): p. 1-14.
34. de Vries, R.P., et al., *Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus Aspergillus.* **Genome Biology**, 2017. 18(1): p. 1-45.
35. Benoit, I., et al., *Closely related fungi employ diverse enzymatic strategies to degrade plant biomass.* **Biotechnology for Biofuels**, 2015. 8: p. 1-14.
36. Mäkelä, M.R., et al., *Genomic and exoproteomic diversity in plant biomass degradation approaches among Aspergilli.* **Studies in Mycology**, 2018. 91(1): p. 79-99.
37. Espagne, E., et al., *The genome sequence of the model ascomycete fungus Podospora anserina.* **Genome Biology**, 2008. 9: p. 1-22.
38. Li, J., R.P. de Vries, and M. Peng, *Bioinformatics approaches for fungal biotechnology.* **Encyclopedia of Mycology**, 2020.
39. Delmas, S., et al., *Uncovering the genome-wide transcriptional responses of the filamentous fungus Aspergillus niger to lignocellulose using RNA sequencing.* **PLoS Genetics**, 2012. 8(8): p. e1002875.
40. Saykhedkar, S., et al., *A time course analysis of the extracellular proteome of Aspergillus nidulans growing on sorghum stover.* **Biotechnology for Biofuels**, 2012. 5(1): p. 1-17.

41. Pullan, S.T., et al., *RNA-sequencing reveals the complexities of the transcriptional response to lignocellulosic biofuel substrates in Aspergillus niger.* **Fungal Biology and Biotechnology**, 2014. 1: p. 1-14.
42. Daly, P., et al., *Expression of Aspergillus niger CAZymes is determined by compositional changes in wheat straw generated by hydrothermal or ionic liquid pretreatments.* **Biotechnology for Biofuels**, 2017. 10(1): p. 1-19.
43. Häkkinen, M., et al., *Re-annotation of the CAZy genes of Trichoderma reesei and transcription in the presence of lignocellulosic substrates.* **Microbial Cell Factories**, 2012. 11: p. 1-26.
44. Tian, C., et al., *Systems analysis of plant cell wall degradation by the model filamentous fungus Neurospora crassa.* **Proceedings of the National Academy of Sciences**, 2009. 106(52): p. 22157-22162.
45. Kolbusz, M.A., et al., *Transcriptome and exoproteome analysis of utilization of plant-derived biomass by Myceliophthora thermophila.* **Fungal Genetics and Biology**, 2014. 72: p. 10-20.
46. Daly, P., et al., *Transcriptomic responses of mixed cultures of ascomycete fungi to lignocellulose using dual RNA-seq reveal inter-species antagonism and limited beneficial effects on CAZyme expression.* **Fungal Genetics and Biology**, 2017. 102: p. 4-21.
47. Zapalska-Sozoniuk, M., et al., *Is it useful to use several "omics" for obtaining valuable results?* **Molecular Biology Reports**, 2019. 46: p. 3597-3606.
48. Evans, T.G., *Considerations for the use of transcriptomics in identifying the 'genes that matter' for environmental adaptation.* **The Journal of Experimental Biology**, 2015. 218(12): p. 1925-1935.
49. Adav, S.S., A. Ravindran, and S.K. Sze, *Study of Phanerochaete chrysosporium secretome revealed protein glycosylation as a substrate-dependent post-translational modification.* **Journal of Proteome Research**, 2014. 13(10): p. 4272-4280.
50. Zhang, R., et al., *Fungal cellulases: protein engineering and post-translational modifications.* **Applied Microbiology and Biotechnology**, 2022. 106(1): p. 1-24.
51. Xiong, Y., et al., *The proteome and phosphoproteome of Neurospora crassa in response to cellulose, sucrose and carbon starvation.* **Fungal Genetics and Biology**, 2014. 72: p. 21-33.
52. Beck, M., M. Claassen, and R. Aebersold, *Comprehensive proteomics.* **Current Opinion in Biotechnology**, 2011. 22(1): p. 3-8.
53. Manadas, B., et al., *Peptide fractionation in proteomics approaches.* **Expert Review of Proteomics**, 2010. 7(5): p. 655-663.
54. van Wijk, K.J., *Challenges and prospects of plant proteomics.* **Plant Physiology**, 2001. 126(2): p. 501-508.
55. van der Werf, M.J., et al., *Microbial metabolomics: toward a platform with full*

*metabolome coverage.* **Analytical Biochemistry**, 2007. 370(1): p. 17-25.

56. Bartel, J., J. Krumsiek, and F.J. Theis, *Statistical methods for the analysis of high-throughput metabolomics data.* **Computational and Structural Biotechnology Journal**, 2013. 4(5): p. e201301009.
57. Rocca-Serra, P., et al., *Data standards can boost metabolomics research, and if there is a will, there is a way.* **Metabolomics**, 2016. 12: p. 1-13.
58. Fekete, E., et al., *D-Galactose uptake is nonfunctional in the conidiospores of Aspergillus niger.* **FEMS Microbiology Letters**, 2012. 329(2): p. 198-203.
59. Meijer, M., et al., *Growth and hydrolase profiles can be used as characteristics to distinguish Aspergillus niger and other black Aspergilli.* **Studies in Mycology**, 2011. 69(1): p. 19-30.
60. Amselem, J., et al., *Genomic analysis of the necrotrophic fungal pathogens Sclerotinia sclerotiorum and Botrytis cinerea.* **PLoS Genetics**, 2011. 7(8): p. e1002230.
61. Floudas, D., et al., *The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes.* **Science**, 2012. 336(6089): p. 1715-1719.
62. De Wit, P.J., et al., *The genomes of the fungal plant pathogens Cladosporium fulvum and Dothistroma septosporum reveal adaptation to different hosts and lifestyles but also signatures of common ancestry.* **PLoS Genetics**, 2012. 8(11): p. e1003088.
63. Berka, R.M., et al., *Comparative genomic analysis of the thermophilic biomass-degrading fungi Myceliophthora thermophila and Thielavia terrestris.* **Nature Biotechnology**, 2011. 29(10): p. 922-927.
64. Klaubauf, S., et al., *Similar is not the same: differences in the function of the (hemi-) cellulolytic regulator XlnR (Xlr1/Xyr1) in filamentous fungi.* **Fungal Genetics and Biology**, 2014. 72: p. 73-81.
65. Benocci, T., et al., *Deletion of either the regulatory gene ara1 or metabolic gene xki1 in Trichoderma reesei leads to increased CAZyme gene expression on crude plant biomass.* **Biotechnology for Biofuels**, 2019. 12(1): p. 1-20.
66. Garrigues, S., et al., *The cultivation method affects the transcriptomic response of Aspergillus niger to growth on sugar beet pulp.* **Microbiology Spectrum**, 2021. 9(1): p. e01064-21.
67. Mäkelä, M.R., et al., *The fungus Aspergillus niger consumes sugars in a sequential manner that is not mediated by the carbon catabolite repressor CreA.* **Scientific Reports**, 2018. 8(1): p. 6655.
68. Mohanta, T.K. and A. Al-Harrasi, *Fungal genomes: suffering with functional annotation errors.* **IMA Fungus**, 2021. 12: p. 1-7.
69. Aguilar-Pontes, M.V., et al., *The gold-standard genome of Aspergillus niger NRRL 3 enables a detailed view of the diversity of sugar catabolism in fungi.* **Studies in Mycology**, 2018. 91: p. 61-78.

70. Raulo, R., M. Kokolski, and D.B. Archer, *The roles of the zinc finger transcription factors XlnR, ClrA and ClrB in the breakdown of lignocellulose by Aspergillus niger.* **Amb Express**, 2016. 6: p. 1-12.

71. Craig, J.P., et al., *Direct target network of the Neurospora crassa plant cell wall deconstruction regulators CLR-1, CLR-2, and XLR-1.* **mBio**, 2015. 6(5): p. e01452-15.

## Supplementary materials

**Table S1.** Carbohydrate metabolism of *A. niger*, *A. nidulans*, *T. reesei* and *D. squalens* described in the KEGG database. ***Available upon request from the author***

**Table S2.** Genome quality assessment results of 200 fungi described in Chapter 6 of this thesis. The highlighted in blue are multiple strains of one species. ***Available upon request from the author***

**Table S3.** Growth profiles of 78 fungi described in Chapter 6 of this thesis. ***Available upon request from the author***

**Table S4.** Summary of CAZy genes, sugar metabolic genes and sugar transporters of 19 fungi described in Chapter 6 of this thesis. ***Available upon request from the author***

**Table S5**. Growth profiles of 19 fungi described in Chapter 6 of this thesis. ***Available upon request from the author***

# Appendix

## Fungi use highly diverse approaches for plant biomass conversion as revealed through bioinformatic analysis

Filamentous fungi play a crucial role as highly efficient microorganisms for the conversion of plant biomass. Their ability to break down complex plant polysaccharides, such as cellulose and hemicellulose, into simple sugars is essential for the recycling of organic material in ecosystems. When it comes to the conversion of plant biomass, two crucial aspects of plant biomass conversion are of paramount importance: primary sugar metabolism and the extensive repertoire of Carbohydrate-Active enZymes (CAZymes) involved in the degradation of complex plant biomass substrates. Therefore, this thesis enhanced our comprehension of the diversity of primary carbon metabolism and the enzymatic capabilities of fungi. By investigating these aspects, we aimed to unravel the intricate mechanisms underlying fungal biomass conversion and shed light on the evolutionary adaptations and functional variations across fungi.

In **Chapter 1**, the background and objectives of this PhD project are described. The first part of the general introduction starts with the composition and conversion of plant biomass. Subsequently, the structure and degradation mechanisms of individual plant polysaccharides are presented, together with the introduction of intracellular metabolic pathways involved in the conversion process of plant biomass. In the second part, a short introduction of basic bioinformatics applications in fungal research is provided. This is extended in **Chapter 2**, which includes an overview and comparison of various bioinformatics tools and methods employed in fungal genomics, transcriptomics, proteomics, and metabolomics studies. This chapter aimed to highlight the importance of utilizing bioinformatics approaches to analyze large-scale datasets and extract meaningful insights from fungal omics data. By leveraging transcriptomic, proteomic, and metabolomic, these multi-omics analyses enable a comprehensive exploration of the diverse strategies employed by fungi in degrading plant biomass.

Knowledge on sugar catabolism is essential to understand the role and abilities of different fungi in natural habitats, but also for their potential as cell factories for the production of biochemicals. In **Chapter 3**, I generated sugar catabolism genetic networks for five fungi, at different taxonomic distances from *Aspergillus niger* by using an orthology-based approach. The findings from my study demonstrated that the reliability of the sugar metabolic model was relatively high in closely related species. Specifically, the sugar catabolic pathways were found to be highly conserved in two other studied Eurotiomycetes, *Aspergillus nidulans* and *Penicillium subrubescens*. However, as the taxonomic distance increased, particularly in the case of the Sordariomycete *Trichoderma reesei* and two Basidiomycota species, the level of diversity in the sugar catabolic pathways significantly increased. This suggests that the conservation of sugar catabolic pathways varies among fungal species, with closer relatives exhibiting higher conservation compared to more distantly related species. In addition, omics-analysis on nine relevant carbon sources were used to compare the response in different fungi, which confirmed the diversity of fungal sugar conversion.

In **Chapter 4**, I investigated the genomic and transcriptomic variations of CAZymes involved in plant biomass degradation in six taxonomic distant fungi (the same as described in **Chapter 3**). Considerable genomic variation and remarkable transcriptomic diversity of CAZymes were revealed for these fungi, demonstrating the large variation in their approach to degrading plant biomass. In addition, the specific carbon utilization ability inferred from genomics and transcriptomics was compared with the fungal growth profile on corresponding sugars. This comparison aimed to enhance our understanding of the conversion process and shed light on the relationship between gene expression and functional outcomes in terms of fungal growth on specific carbon sources. By integrating genomic, transcriptomic, and phenotypic data, we gained novel insights into the complex interplay between gene regulation, enzyme activity, and carbon utilization in plant biomass degradation.

In **Chapter 5**, I explored the diversity of fungi in their approach to converting crude plant biomass. Through a comparative analysis in five fungi, I examined the gene expression patterns of CAZymes, sugar transporters, and sugar metabolic genes. Notably, we observed significant differences in these profiles that were influenced by the specific time point, substrate, and fungal species. These findings indicate the presence of strong time-, substrate-, and species-specific gene regulation during fungal adaptation to crude plant biomass. Additionally, I expanded our analysis by comparing the transcriptome profiles on crude plant biomass to such profiles on various plant-derived monosaccharides. This comprehensive approach allowed us to uncover complex gene regulation patterns and substrate preferences exhibited by these fungi during their adaptation to crude plant biomass.

Taken together, the results described in this thesis enriched our knowledge on the diversity of plant biomass conversion strategies of fungi. In the final **Chapter 6**, I address several important points regarding the comparison between our sugar metabolic models and the KEGG database. In addition, I discuss the challenges associated with the separation of enzymatic functions by using phylogenetic trees. Also considerations of using multi-omics data were presented. Moreover, I presented some ideas for possible future research directions based on the extended results described in this thesis.

## Schimmels gebruiken zeer diverse strategieën voor de conversie van plantenbiomassa, ontrafelt door bioinformatische analyse

Filemanteuze schimmels spelen een cruciale rol als zeer efficiënte micro-organismen voor de omzetting van plantenbiomassa. Hun vermogen om complexe plant polysachariden, zoals cellulose en hemicellulose, af te breken tot eenvoudige suikers, is essentieel voor de recycling van organisch materiaal in ecosystemen. Als het gaat om de omzetting van plantenbiomassa, zijn twee cruciale aspecten van de omzetting van plantenbiomassa van het grootste belang: het primaire suikermetabolisme en het uitgebreide repertoire van koolhydraat-actieve enzymen (CAZymes) die betrokken zijn bij de afbraak van complexe plantenbiomassa substraten. Dit proefschrift heeft ons begrip van de diversiteit van het primaire koolstofmetabolisme en de enzymatische mogelijkheden van schimmels vergroot. Door deze aspecten te onderzoeken, wilden we de ingewikkelde mechanismen die ten grondslag liggen aan de conversie van schimmelbiomassa ontrafelen en licht werpen op de evolutionaire aanpassingen en functionele variaties tussen schimmels.

In **Hoofdstuk 1** worden de achtergrond en doelstellingen van dit promotieonderzoek beschreven. Het eerste deel van de algemene inleiding behandelt de samenstelling en omzetting van plantenbiomassa. Vervolgens worden de structuur en afbraakmechanismen van individuele plant polysachariden gepresenteerd, samen met de introductie van intracellulaire metabole routes die betrokken zijn bij het conversieproces van plantenbiomassa. In het tweede deel wordt een korte inleiding gegeven van basale bioinformatica toepassingen in het schimmelonderzoek. Dit wordt uitgebreid in **Hoofdstuk 2**, dat een overzicht en vergelijking bevat van verschillende bioinformatica tools en methoden die worden gebruikt in genomische, transcriptomische, proteomische en metabolomische onderzoek van schimmels. Dit hoofdstuk was bedoeld om het belang te benadrukken van het gebruik van bioinformatica benaderingen om grootschalige datasets te analyseren en zinvolle inzichten

te extraheren uit omics gegevens van schimmels. Door gebruik te maken van transcriptomics, proteomics en metabolomics, maken deze multi-omics analyses een uitgebreide verkenning mogelijk van de diverse strategieën die schimmels gebruiken bij het afbreken van plantenbiomassa.

Kennis over het suiker katabolisme is essentieel om de rol en capaciteiten van verschillende schimmels in natuurlijke habitats te begrijpen, maar ook voor hun potentieel als celfabrieken voor de productie van biochemicaliën. In **Hoofdstuk 3** heb ik genetische netwerken voor suikerkatabolisme gegenereerd voor vijf schimmels, op verschillende taxonomische afstanden van *Aspergillus niger*, met behulp van een op orthologie gebaseerde benadering. De bevindingen van mijn onderzoek toonden aan dat de betrouwbaarheid van het suiker metabolisme model relatief hoog was bij nauw verwante soorten. In het bijzonder bleken de suiker katabole routes sterk geconserveerd te zijn in twee andere bestudeerde Eurotiomyceten, *Aspergillus nidulans* en *Penicillium subrubescens*. Naarmate de taxonomische afstand echter toenam, met name in het geval van de Sordariomycete *Trichoderma reesei* en twee Basidiomycota soorten, nam de diversiteit in de katabolische routes van suiker aanzienlijk toe. Dit suggereert dat het behoud van suiker katabole routes varieert tussen schimmelsoorten, waarbij nauwere verwanten een hogere conservering vertonen in vergelijking met verder weg verwante soorten. Daarnaast werd omics analyse van negen relevante koolstofbronnen gebruikt om de respons in verschillende schimmels te vergelijken, wat de diversiteit van de suikerconversie van schimmels aantoonde.

In **Hoofdstuk 4** onderzocht ik de genomische en transcriptomische variatie van CAZymes die betrokken zijn bij de afbraak van plantenbiomassa in zes taxonomisch diverse schimmels (dezelfde als beschreven in **Hoofdstuk 3**). Aanzienlijke genomische variatie en opmerkelijke transcriptomische diversiteit van CAZymes werd onthuld voor deze schimmels, wat de grote variatie aantoont in hun benadering om plantenbiomassa af te breken. Bovendien werd het specifieke vermogen voor het gebruik van verschillende

koolstoffen afgeleid uit genomische en transcriptomische data vergeleken met het schimmelgroeiprofiel op overeenkomstige suikers. Deze vergelijking was bedoeld om ons begrip van het conversieproces te vergroten en licht te werpen op de relatie tussen genexpressie en functionele resultaten in termen van schimmelgroei op specifieke koolstofbronnen. Door genomische, transcriptomische en fenotypische gegevens te integreren, hebben we nieuwe inzichten verkregen in het complexe samenspel tussen genregulatie, enzymactiviteit en koolstofbron gebruik bij de afbraak van plantenbiomassa.

In **Hoofdstuk 5** onderzocht ik de diversiteit van schimmels in hun strategie voor het omzetten van ruwe plantaardige biomassa. Door middel van een vergelijkende analyse in vijf schimmels onderzocht ik de genexpressiepatronen van CAZymes, suiker transporters en suiker metabole genen. We hebben met name significante verschillen waargenomen in deze profielen die werden beïnvloed door het specifieke tijdstip, het substraat en de schimmel soort. Deze bevindingen duiden op de aanwezigheid van sterke tijd-, substraat- en soortspecifieke genregulatie tijdens de aanpassing van schimmels aan ruwe plantenbiomassa. Daarnaast heb ik onze analyse uitgebreid door de transcriptomische profielen op ruwe plantenbiomassa te vergelijken met dergelijke profielen op verschillende van planten afgeleide monosacchariden. Deze alomvattende aanpak stelde ons in staat om complexe genregulatiepatronen en substraatvoorkeuren van deze schimmels bloot te leggen tijdens hun aanpassing aan ruwe plantenbiomassa.

Alles bij elkaar genomen, verrijkten de resultaten beschreven in dit proefschrift onze kennis over de diversiteit van strategieën voor het omzetten van plantenbiomassa door schimmels. In het laatste **Hoofdstuk 6** behandel ik verschillende belangrijke punten met betrekking tot de vergelijking tussen onze suiker metabole modellen en de KEGG-database. Daarnaast bespreek ik de uitdagingen die samenhangen met de scheiding van enzymatische functies door middel van fylogenetische bomen. Ook werden overwegingen voor het gebruik van multi-omics data gepresenteerd. Bovendien presenteerde ik

enkele ideeën voor mogelijke toekomstige onderzoeksrichtingen op basis van de uitgebreide resultaten die in dit proefschrift beschreven zijn.

Jiajia Li was born on January 1st, 1992 in Shanxi, China. She enrolled in the Biological Engineering bachelor program of Dalian University in 2011 and graduated in 2015. In September 2015, she continued with her Master study in Northwest A&F University. She did her MSc project in the State Key Laboratory of Crop Stress Biology for Arid Areas under the supervision of Prof. dr. Chuang Ma and received her master degree in 2018. In July of the same year, she joined the Medical Division of Biomarker Technologies Company (http://www.biomarker.com.cn/) as a Bioinformatics R&D Engineer. In October 2019, she registered as a PhD student at Utrecht University, and started her PhD research in the Fungal Physiology group at the Westerdijk Fungal Biodiversity Institute (Utrecht, The Netherlands), under the supervision of Prof. dr. ir. Ronald P. de Vries, Prof. dr. Berend Snel and Dr. Mao Peng.

## List of publications

**Li, J.**, de Vries, R.P. and Peng, M., 2020. Bioinformatics Approaches for Fungal Biotechnology. In Encyclopedia of Mycology. Elsevier BV.

**Li, J.**, Chroumpi, T., Garrigues, S., Kun, R. S., Meng, J., Salazar-Cerezo, S., ... and de Vries, R. P., 2022. The Sugar Metabolic Model of Aspergillus niger Can Only Be Reliably Transferred to Fungi of Its Phylum. Journal of Fungi, 8(12), 1315.

**Li, J.**, Wiebenga, A., Lipzen, A., Ng, V., Tejomurthula, S., Zhang, Y., Grigoriev, I.V., Peng, M. and de Vries, R.P., 2023. Comparative genomics and transcriptomics analyses reveal divergent plant biomass-degrading strategies in fungi. Journal of Fungi, 9(8), 860.

As I reflect on my PhD journey and write this final chapter of "Acknowledgements," I can't help but feel a mix of emotions. I expected a sense of relief, but instead I find myself unable to write down any words for a long time. Looking back, my PhD journey has been filled with moments of opportunities and challenges, shaping me into the person I am today. I can wholeheartedly say that I have cherished every step of this journey, and it will forever hold a special place in my heart as the most important and cherished memory of my study life.

Throughout these four years, I have been immensely fortunate to receive help, support, and care from many people, each contributing in their unique way. I am extremely grateful to my supervisors, colleagues, friends, and family who have played an important role in my study and life, providing unwavering support, encouragement, and guidance. I cannot express how much I have learned, grown and enjoyed working with all of you! But what I can say for sure is that without you, I could not have made it. To all of you, please accept my heartfelt thanks!!!

First and foremost, I would like to thank my major supervisor and promoter, Ronald de Vries. I really really really appreciate your help. I remembered that after I failed to apply for the China Scholarship Council (CSC) grant the first time, you were still willing to encourage me and support my application for the CSC grant again. When I joined our Fungal Physiology Group, my English was limited, but you never made me feel inadequate. Instead, you patiently guided me and gave me unwavering encouragement. As time went on, I noticed a gradual change within myself. With your constant support and guidance, I started gaining confidence and expressing my ideas more freely. I feel incredibly happy and fortunate to be one of your PhD students. Your mentorship has been invaluable to me, and I am grateful for all the knowledge and growth you have instilled in me. Thank you for believing in me and helping me achieve what I have today. I have to say: You are such an amazing supervisor! During these four years, when I encountered questions and

struggled with experimental results, every time I turned to you for guidance and shared my concerns, you never failed to provide insightful solutions and guide me back on the right path. Especially in this year, I am pregnant. Your timely and effective help and guidance, as well as your understanding, have enabled me to successfully complete my thesis during this challenging period. I really appreciate it. Not only in terms of the study, but you also give us a lot of cares in life. Your efforts to organize various activities, such as dinners, walking, BBQ, and horse riding, have not only helped us fit into the group but also provided us with opportunities to practice our English and feel more at ease. What's more, you send everyone New Year's cards and birthday wishes every year, which is so warm. Thank you again for everything you have done, it has made my life in the Netherlands beautiful and fulfilling!

I would also like to thank my other supervisor and promotor, Berend Snel. I am so happy that you are involved in my project, providing me with different ways of thinking and designing the research. Every time when Ronald and me were not sure about some ideas, we concluded as "let's have a meeting with Berend", leading to fruitful discussions. You helped me with focusing on the main research questions. Your critical thinking, valuable suggestions and constructive feedback played an important role in my PhD research. Especially during the writing phase, your advice and comments significantly improved the structure, logic, and readability of my manuscripts. Both you and Ronald are amazing supervisors, and I thoroughly enjoyed collaborating with you.

Special and great thanks to my co-promotor from Westerdijk Institute, Mao Peng. You are one of the key people who helped me the most throughout my whole research (you have contributed to every chapter). It was a great pleasure to have your guidance during the four years of my PhD studies. You are not only a great daily supervisor but also a wonderful friend. You are supportive and helpful, both in my study and life. From the moment I joined our group, you helped me get familiar with the lab environment and guided

me to understand the projects. You are always so patient. I am truly grateful for your help.

I also want to thank Miia Mäkelä from Helsinki University. Although we have only met a few times, I really appreciate your help, especially with my manuscripts, as you always revised my manuscripts with great patience, and provided many comments and suggestions on my work.

I am also grateful to my defense committee members, Prof. dr. A.F.J.M. van den Ackerveken (Utrecht University), Prof. dr. B. Teusink (Vrije Universiteit Amsterdam), Prof. dr. A. Tsang (Concordia University), Prof. dr. ir. M.W. Fraaije (University of Groningen) and Dr. I. Druzhinina (Royal Botanic Gardens, Kew). Thank you all for taking your time to attend my defense and, most importantly, for your reading and evaluating my PhD thesis.

And now it's time to express my deepest gratitude and thanks to the Westerdijk Institute and entire FP group. It is so great to work here. I not only enjoyed the scientific atmosphere and research environment, but also all the lab-out activities, and coffee/lunch breaks.

To the Chinese girls in D-wing, I would like to say, I am very pleased to meet you and we really had a great time together. 首先是Xinxin（新新）和Jiali（佳丽）（虽然你们已经毕业了），和你们在一起的三年，真的很开心（逛街，聊八卦，哈哈哈）。不过作为你们的师妹，其实是倍感压力的。你们在博士期间，科研都做的那么好。现在你们都在自己的工作岗位上发光发热，我真心祝福你们一直顺顺利利。Dujuan（杜娟），你总是那么开朗和乐观，你的活泼可以感染每一个接触你的人。其次我想说做你邻居一年多，真的无比开心和幸福。记得我感染新冠的那一次，你做饭给我吃，饭菜特别可口，我特别感动。之后还发生了好多事情，比如我好几次被锁在门外，我们一起去游泳，一起和外国小伙伴们玩耍，孕期你陪我散步等等，总之都是美好的回忆。Li Xu（许力），你很细心和踏实。和你在一起的时候，总是能被你照顾到，无论是生活还是工作。希望你在最后的一年里，收获累累的实验结果，多发表文章，多出去玩一玩，好好享受美好的生活。Lin Zhao（赵琳），美丽温柔的小妹子，工作那么那么认真，希望你最后一年也收获满满。希

望你们三个都顺顺利利，明年毕业找到好工作。Xin Zhang（张鑫），你是特别温暖的人。感谢你在这最后阶段一直给我加油打气，希望你很快找到心仪的工作，顺顺利利，开开心心的。Yanfang（艳芳），特别喜欢和你聊天，很放松，而且你的蛋糕做的真的无比好吃。感恩遇见你们！

Mar, Ola, Alessia and Raquel, thanks for being my best friends and colleagues in the Netherlands. You are all such nice and warm people. I truly enjoyed the moments we spent chatting, having fun, and dining together. Especially after I was pregnant, you always cared about me, helped me, and supported me. Thank you for being so supportive and helpful whenever I needed! I cannot thank you enough for being such caring and supportive friends. I also would like to thank Agata and Astrid. Even though we don't have much experimental interaction, you are always warm and friendly when I need your help. The best of luck with your PhD and your future job.

During these four years, I had the pleasure to meet and work with many people in the Westerdijk Institute. To Pedro (our former director), Manon, Sandra, Adiphol, Roland, Tania, Ronnie, Sumitha, Chendo, Natalia…Thank you all for being so helpful and friendly, making the Westerdijk Institute such a great place to be.

Of course, I would like to give my heartfelt thanks and sincere love to my family in China! 爸爸妈妈，爷爷奶奶，弟弟妹妹，我想感谢你们对我无尽的爱，理解和包容。无论什么时候，我总能得到你们无条件的支持和肯定。这让我特别安心和踏实，可以快乐地做我自己。谢谢你们，永远爱你们！还有我的姑姑们，谢谢你们对我一直以来的关心和记挂。Next, my greatest thanks to my husband, Tingchao, and my daughter, Aili. Thank you for being in my life. 毛廷超，我的“猪队友”，时常觉得你笨的不行，却也是你陪我走了这么远，为我遮风挡雨，希望未来的我们越来越好，携手共度余生。小艾荔，我的小宝贝，感谢你的出现，虽然在这样一个紧张的阶段。妈妈对你又愧疚又惊喜，愧疚的是你陪着妈妈加了好多班，担心焦虑了很多事情；惊喜你的出现，期待你从那么个小不点一天天变大，感恩！希望你健健康康地成长，爸爸

妈妈会一直爱你，陪伴你。惟愿我儿愚且鲁，无灾无难到公卿。

Last but not the least, I would like to thank the China Scholarship Council (CSC) for providing the financial support that enabled me to pursue my dreams of studying abroad. I'm so lucky to be here and be who I am now.

# THE END