



HAL
open science

From the genome to the transcriptome for the characterization of networks controlling the expression of hydrolytic enzymes in a fungus of industrial interest.

Agustina Llanos

► To cite this version:

Agustina Llanos. From the genome to the transcriptome for the characterization of networks controlling the expression of hydrolytic enzymes in a fungus of industrial interest.. Biotechnology. INSA de Toulouse, 2014. English. NNT : 2014ISAT0029 . tel-02917969

HAL Id: tel-02917969

<https://theses.hal.science/tel-02917969>

Submitted on 20 Aug 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National des Sciences Appliquées de Toulouse (INSA de Toulouse)

Présentée et soutenue par :

Agustina LLANOS

le mercredi 24 septembre 2014

Titre :

From the genome to the transcriptome for the characterization of networks controlling the expression of hydrolytic enzymes in a fungus of industrial interest

École doctorale et discipline ou spécialité :

ED SEVAB : Ingénieries microbienne et enzymatique

Unité de recherche :

Laboratoire d'Ingénierie des Systèmes Biologiques et des Procédés (EAD5)

Directeur/trice(s) de Thèse :

Jean-Luc PARROU

Virginie NEUGNOT-ROUX

Jury :

Jean Marie FRANÇOIS, INSA de Toulouse, France

Arthur RAM, Leiden University, The Netherlands

Thierry ROUXEL, INRA Bioger, France

Markku SALOHEIMO, VTT Technical Research Centre, Finland

David ARCHER, University of Nottingham, UK

Thesis submitted for the Doctoral degree
Université de Toulouse

**From the genome to the transcriptome for the
characterization of networks controlling the
expression of hydrolytic enzymes in a fungus of
industrial interest**

Agustina LLANOS
Supervision: Jean-Luc PARROU
Co-supervision: Virginie Neugnot-Roux

July 2014

Abstract

Talaromyces versatilis is an industrially important enzymes producing filamentous fungus. Adisseo Company commercializes the enzymatic cocktail, produced from *T. versatilis* fermentation, with the name of Rovabio™. This cocktail is applied as an animal feed additive as it contains a wide variety of hydrolytic enzymes that can degrade the polysaccharides present in the seed-coat and thus improves the digestibility and increases the nutritional value of the agricultural raw materials. Although efforts have been done to study different aspects of the biology of *T. versatilis*, very little is known about this fungus. This study aimed to describe the regulatory networks of genes encoding plant cell wall-degrading enzymes from this biotechnologically important fungus using genomic and transcriptomic approaches.

Having a correct annotation of the genomic sequence together with efficient tools for genome engineering are essential for downstream functional genomics works and characterization of the regulatory networks. Therefore, the first task carried out an analysis of the genomic sequence and a manual curation of the annotation, which led us to assess the vast genetic potential of *T. versatilis* for the production and secretion of hydrolytic enzymes involved in the degradation of lignocellulosic materials. Secondly, I adapted a gene deletion system initially designed for *Aspergillus niger*. This method allows recycling of the selection marker and is efficient in a non-homologous end-joining (NHEJ)-proficient strain (Delmas, Llanos et al., 2014, AEM). During this work, two deletion mutants of *T. versatilis* were obtained: $\Delta xlnR$ and $\Delta clrA$.

Towards better understanding of the regulatory network, I first contributed to an RNAseq-based transcriptomic study that was performed on the wild type strain of *T. versatilis* exposed to glucose and wheat straw as carbon sources. The data showed a massive increase in transcript levels of numerous genes, in particular those encoding hydrolytic enzymes, when the mycelium was incubated with lignocellulose.

If RT-qPCR is indeed a suitable technique to study a limited number of genes in a large variety of conditions, data normalisation is a critical step of the workflow that can lead to incorrect biological interpretation of gene regulation. The work done on the RNA-seq data led me to reconsider the use of the classical reference genes, since most of them exhibited expression changes in the presence of lignocellulosic substrate. I therefore identified a new set of putative reference genes and validated their expression stability by RT-qPCR in *T. versatilis* cultivated under more than 30 different conditions. Then, I collected about a hundred RNA-seq datasets from 18 phylogenetically distant filamentous fungi, to demonstrate that the use of the suitable candidates for RT-qPCR data normalisation in *T. versatilis* can be extended to other fungi (Llanos et al., 2014 BMC genomics (minor revisions)). Thereafter, I performed a more detailed RT-qPCR based transcriptional study of a group of genes of interest, in a wide variety of conditions and in 2 strains, the wild-type and the $\Delta xlnR$ mutant. The analysis of expression data of the genes of interest allowed to identify genes with similar expression patterns, which probably share the same regulatory mechanisms and also the substrates that act as inducers for their expression.

Résumé

Talaromyces versatilis est un champignon filamenteux d'intérêt industriel grâce à sa capacité de production d'enzymes hydrolytiques. La Société Adisseo commercialise un cocktail enzymatique produit par fermentation à partir de *T. versatilis*, sous le nom de Rovabio™. Ce cocktail est utilisé en tant qu'additif alimentaire en nutrition animale, car la grande variété d'enzymes hydrolytiques qu'il contient peut dégrader les polysaccharides présents dans l'enveloppe des céréales, améliorant ainsi la digestibilité la valeur nutritionnelle des matières premières agricoles. Malgré les efforts consentis pour mieux connaître la biologie de *T. versatilis*, très peu est connu sur ce champignon. L'étude présentée ici vise à décrire les réseaux de régulation qui contrôlent l'expression des gènes codant pour ces enzymes hydrolytiques, en utilisant des approches génomiques et transcriptomiques.

Avoir accès à une annotation correcte de la séquence génomique et posséder les outils nécessaires pour l'ingénierie génétique sont essentiels pour réaliser des études de génomique fonctionnelle. Donc, le premier volet de cette thèse a été l'analyse de la séquence génomique et la curation manuelle de l'annotation, ce qui nous a conduits à évaluer le vaste potentiel génétique de *T. versatilis* pour la production et la sécrétion d'enzymes hydrolytiques impliquées dans la dégradation de la lignocellulose. Deuxièmement, un système de délétion des gènes initialement conçu pour *Aspergillus niger* a été adapté à *T. versatilis*. Cette méthode permet le recyclage du marqueur de sélection et est efficace dans des souches dont le système NHEJ est actif (Delmas, *et al.*, 2014, AEM). Au cours de ce travail, deux mutants de délétion de *T. versatilis* ont été obtenus: $\Delta xlnR$ et $\Delta clrA$.

La première approche mise en place pour avoir une meilleure compréhension des réseaux de régulation via une vue globale du transcriptome, fut l'utilisation de la technique de RNAseq sur trois échantillons issus de la souche sauvage de *T. versatilis* exposée au glucose, à la paille de blé et au glucose et paille de blé simultanément comme sources de carbone, respectivement. Les données ont montré une augmentation massive des niveaux d'expression de nombreux gènes, en particulier ceux codant pour des enzymes hydrolytiques, lorsque le mycélium est exposé à la lignocellulose.

Enfin, la dernière partie du projet s'est appuyée sur la RT-qPCR, technique appropriée pour étudier un nombre limité de gènes dans une grande variété de conditions. Toutefois la normalisation des données est une étape essentielle du flux de travail qui peut conduire à une interprétation biologique incorrecte de la régulation des gènes. Le travail effectué sur les données de RNAseq nous a amené à reconsidérer la nature des gènes de référence classiquement utilisés, puisque la plupart d'entre eux présentaient des changements d'expression considérables en présence de lignocellulose. En conséquence, un nouvel ensemble de gènes de référence putatifs a été identifié et la stabilité de leur expression validée par RT-qPCR chez *T. versatilis* cultivé dans plus de 30 conditions différentes. Des jeux de données de RNAseq de 18 champignons filamenteux phylogénétiquement éloignés ont par ailleurs été collectés, afin de démontrer que la sélection des gènes candidats pour la normalisation des données de RT-qPCR chez *T. versatilis* peut être étendue à d'autres champignons (Llanos *et al.*, 2014, BMC Genomics). Ces aspects méthodologiques validés, nous avons enfin réalisé une étude plus détaillée de la transcription d'un groupe de gènes d'intérêt par RT-qPCR, dans une grande variété de conditions et 2 souches différentes, la souche sauvage et la mutante $\Delta xlnR$. L'analyse de ces données a permis d'identifier des gènes aux profils d'expression similaires, qui répondent de la même façon aux substrats inducteurs et qui partagent probablement les mêmes mécanismes de régulation.

Publications

Llanos A, François JM and Parrou JL. Tracking the best reference genes for RT-qPCR data normalization in filamentous fungi. BMC Genomics. Submitted in April 2014.

Delmas S, **Llanos A**, Parrou JL, Kokolski M, Pullan S, Shunburne L, Archer D. Development of an Unmarked Gene Deletion System for the Filamentous Fungi *Aspergillus niger* and *Talaromyces versatilis*. Applied and Environmental Microbiology. Accepted in March 2014.

Acknowledgements

The following pages are merely a résumé of the experiences of the last three years, not only in the lab, but also in life... they could never have been possible without the support of all the people around me.

I will start by thanking Nic Lindley for giving me the opportunity to work in the LISBP.

Thanks to Jean Marie François for letting me be part of his group, for the opportunity to undertake this project and for his support during the last three years.

I would also like to thank Jean-Luc Parrou and David Archer for their support, wise advice and guidance throughout this work. I specially appreciated working in the 2 labs, Toulouse and Nottingham, and learning two different styles of work. It was a wonderful and enriching experience for me. Thank you both!

I thank Marc Maestracci, Virginie Nugneot-Roux and Olivier Guais for giving me the opportunity to work with you and for the “industrial approach” during this works. It was extremely interesting to know how the research in a private company is carried on. Thanks to all the CINABio team as well!

I am grateful to the ANRT and the Adisseo Company for the founding for this project.

During the first year in Nottingham I enjoyed the company of lots of very nice people. First, Lee Shunburne, Steven Pullan, Stéphane Delmas and Matthew Kokolski, how were extremely helpful and introduced me to the fungi world. There were also “the girls”: Laure, Heather, Kim and Roxane, whose friendship I treasure.

The last two years spent in Toulouse were full of pleasant and funny moments. I would like to thank all the EAD5 team; Mari-Ange, Adilia, Hélène, Gusti, Thomas, Jean-Pascal, Laurent. A special thanks to all my colleagues and friends from the “students’ room”: Ceren, Marion la blonde, Marion la brune, Amélie, Marjorie, Ran, Tian, Romain, Debora, Florence, Luce, Xu...

I thank the people that helped me to carry on some of the experiments and analyse the data (not always easy...), without them it would have been impossible: Martin Blythe, Lidwine Trouilh and Sébastien Déjean. Thank you for your patience and for lending me your expertise.

Finally I would like to thank my family, specially my mum and dad (wherever you are), and friends for giving me their love and support throughout my PhD.

Table of contents

<i>Abstract</i>	3
<i>Résumé</i>	5
<i>Publications</i>	7
<i>Acknowledgements</i>	9
<i>Table of contents</i>	11
<i>Abbreviations</i>	17
<i>Presentation of the manuscript</i>	19
1. General introduction	23
1.1. Fungal biology	23
1.1.1. Fungal growth and development	23
1.1.1.1. Spore formation and dispersal	23
1.1.1.2. Spore germination and vegetative growth.....	25
1.1.1.3. Hyphae and apex structure	26
1.1.1.4. The secretion pathway	28
1.1.2. <i>Talaromyces versatilis</i> in the fungal classification.....	33
1.1.3. Filamentous fungi genomes	34
1.1.4. <i>Talaromyces versatilis</i> genome	37
1.2. Plant polymers degradation by filamentous fungi	38
1.2.1. Industrial application of fungal enzymes.....	38
1.2.2. Genetic potential of filamentous fungi for lignocellulosic material deconstruction	39
1.2.2.1. Plant cell wall composition	39
1.2.2.2. CAZymes	41

1.2.2.2.1.	Cellulases.....	43
1.2.2.2.2.	Hemicellulases	44
1.2.2.2.3.	Ligninases	45
1.2.2.2.4.	Auxiliary activities	45
1.2.3.	Production of hydrolytic enzymes by <i>Talaromyces versatilis</i>	47
1.3.	Regulation of the expression of cellulase and hemicellulase genes in filamentous	
fungi	47
1.3.1.	The XlnR regulator	49
1.3.2.	The AraR regulator	50
1.3.3.	The CLR1 and CLR2 regulators	51
1.3.4.	The Pal pathway	52
1.3.5.	The ACEI and ACEII transcription factors.....	54
1.3.6.	The AreA regulator	55
1.3.7.	The carbon catabolite repression system.....	55
1.3.8.	The Unfolded Protein Response and the HacA transcription factor	57
1.3.9.	Noncoding RNAs and Antisense transcription.....	58
1.4.	Tools for the study of the genome and the transcriptome	61
1.4.1.	Genome sequencing technologies	61
1.4.2.	Reference genes for RT-qPCR data normalisation.....	67
1.4.3.	Molecular tools for the genetic modification of filamentous fungi	69
2.	<i>General materials and methods</i>.....	75
2.1.	Microbiology techniques	75
2.1.1.	Chemicals and Reagents.....	75
2.1.2.	Strains.....	75
2.1.3.	Culture conditions	75
2.1.3.1.	Growth on agar plates.....	77
2.1.3.2.	Growth curves in liquid media.....	77

2.1.4.	Harvest and storage of mycelia	78
2.1.5.	Transformation of <i>Talaromyces versatilis</i>	78
2.1.6.	Hydrolytic activity test by congo red dye	79
2.2.	Basic molecular biology techniques.....	80
2.2.1.	Nucleic acids isolation	80
2.2.1.1.	Genomic DNA extraction from <i>Talaromyces versatilis</i>	80
2.2.1.2.	Rapid gDNA extraction	81
2.2.1.3.	RNA extraction from mycelia.....	81
2.2.1.4.	RNA extraction from conidia and germinating conidia	82
2.2.2.	Polymerase Chain Reaction (PCR)	83
2.2.3.	Plasmid construction	84
2.3.	Gene expression analysis.....	86
2.3.1.	PCR based methods.....	86
2.3.1.1.	Primer design for mRNA quantification.....	86
2.3.1.2.	Reverse-transcription	87
2.3.1.2.	Semi-quantitative PCR.....	88
2.3.1.3.	Quantitative PCR	88
2.3.2.	RNA sequencing method	89
2.3.2.1.	Experimental design	89
2.3.2.2.	Sequencing	89
2.3.2.3.	First analysis of the sequencing results.....	90
2.3.2.4.	Annotation pipeline.....	90
2.3.2.5.	Second analysis of the RNA-seq data	91
2.3.2.6.	Differential gene expression.....	91
2.3.2.7.	GO enrichment analysis.....	92
2.4.	Computational analysis	92
2.4.1.	Statistical analysis.....	92
2.4.2.	Genomic sequence analysis.....	93

3.	<i>Talaromyces versatilis</i> mycelia growth and development.....	95
3.1.	Talaromyces versatilis growth with different carbon sources and comparison to other filamentous fungi	95
3.2.	Growth on liquid media with different N sources	99
3.3.	Discussion	101
4.	<i>Genomic sequence annotation</i>	105
4.1.	Manual curation of the genome.....	105
4.1.1.	Glycoside hydrolases encoding genes	109
4.1.2.	Transcriptional regulators involved in the GH production and secretion	113
4.1.3.	Secretion pathway.....	117
4.2.	Improvement of the genomic annotation using the RNA-seq data	119
4.3.	Discussion	120
5.	<i>Study of the transcriptome of Talaromyces versatilis</i>	127
5.1.	Genome wide transcriptome study using RNA-seq of Talaromyces versatilis exposed to lignocellulose	127
5.1.1.	First RNA-seq results analysis	127
5.1.1.1.	Expression of the glycoside hydrolases genes.....	132
5.1.1.2.	Accessory proteins encoding genes induced by wheat straw	138
5.1.1.3.	Transcription factors	139
5.1.1.4.	Antisense transcription	142
5.1.2.	RNA-seq remapping to the new genome annotation	147
5.2.	Analysis of the expression of the GH-encoding genes by RT-qPCR	153
5.2.1.	Tracking the best reference genes for RT-qPCR data normalisation.....	153
5.2.2.	RT-qPCR analysis of the expression levels of GH-encoding genes over different carbon sources	158

5.2.2.1.	Global analysis of the expression patterns of the hydrolytic enzymes	161
5.2.2.2.	The class 1 expression pattern: The cellulase genes	172
5.2.2.3.	The class 2 expression pattern: The main hemicellulase genes	174
5.2.2.4.	The class 3 expression patter: the minor hemicellulase genes	177
5.2.2.5.	The class 4 expression pattern: the β -xylosidase genes	179
5.2.2.6.	The class 5 expression pattern: the hydrophobin gene	181
5.3.	Discussion	183
6.	<i>Effect of the deletion of key transcription factors over the expression of the GH encoding genes.....</i>	193
6.1.	Construction of the deletion mutants	193
6.1.1.	Construction of a <i>Talaromyces versatilis</i> Δ pyrG strain	193
6.1.2.	Development of an unmarked gene deletion system for <i>Talaromyces versatilis</i>	196
6.2.	Deletion mutants characterisation	200
6.3.	Study of the expression levels of the of genes of interest in a selection of substrates	203
6.4.	Discussion	214
7.	<i>General conclusions and perspectives</i>	219
7.1.	Conclusions	219
7.2.	Perspectives	224
Appendix I.....	227
Appendix II.....	229
Appendix III.....	261
Appendix IV.....	265

Appendix V.....	267
Appendix VI.....	271
Appendix VII.....	273
Appendix VIII.....	281
References	283

Abbreviations

5'FOA	5'fluoroorotic acid
AA	Auxiliary activities
Abf	Arabinofuranosidase
ADP	Adenosine diphosphate
AMM	<i>Aspergillus</i> minimal medium
AMT	<i>Agrobacterium</i> mediated transformation
AraR	Arabinolytic regulator
ATP	Adenosine triphosphate
AVC	Apical vesicle cluster
BLAST	Basic Local Alignment Search Tool
bp	Base-pair
CAZy	Carbohydrate active enzyme
Cbh	Cellobiohydrolase
CBM	Carbohydrate binding module
CCR	Carbon catabolite repression
CD	Catalytic domain
cDNA	Complementary DNA
CE	Carbohydrate esterase
CMC	Carboxymethyl cellulose
CreA	Carbon catabolite repressor
DE	Differentially expressed
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
dsRNA	double strand RNA
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
ER	Endoplasmic reticulum
ERAD	ER-associated protein degradation
FC	Fold change
gDNA	Genomic DNA
GEO	Gene expression omnibus
GH	Glycoside hydrolase
GO	Gene ontology
GOI	Gene of interest
GTF	General transfer format
HAC	Hierarchical ascendant classification
HR	Homologous recombination
HsbA	Hydrophobic surface binding protein A
Hsp70	Heat shock protein 70
IGV	Integrative Genomics Viewer
IMI	International Mycological Institute

IPTG	Isopropyl β -D-1-thiogalactopyranoside
Ire1	Inositol-requiring protein 1
JGI	Joint Genome Institute
LB	Luria-Bertani medium
lncRNA	Long noncoding RNA
ManR	Mannanolytic regulator
Mb	Mega base-pairs
MIQE	Minimum Information for Publication of Quantitative PCR Experiments
mRNA	Messenger RNA
NAT	Natural antisense transcript
NF	Normalisation factor
NHEJ	Non-homologous end joining
ORF	Open reading frame
PBSA	Polybutylene succinate-co-adipate
PCR	Polymerase chain reaction
PDA	Potato dextrose agar
PEG	Polyethylene glycol
PET	poly(ethylene terephthalate)
PL	Polysaccharide lyase
piRNA	Piwi-interacting RNA
RdRP	RNA-dependent RNA polymerase
RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
RNAi	RNA interference
RPKM	Reads per kilobase of exon model per million mapped reads
RPM	Revolutions per minute
rRNA	ribosomal RNA
RT-qPCR	Reverse transcription quantitative PCR
SDS	Sodium dodecyl sulphate
siRNA	Silencing RNA
SRP	Signal recognition particle
ssRNA	Single strand RNA
TF	Transcription factor
TMM	<i>Trichoderma</i> minimal medium
tRNA	Transfer RNA
TvMM	<i>T. versatilis</i> minimal medium
UPR	Unfolded protein response
UV	Ultra violet
XlnR	Xylanolytic regulator
Xyn	Xylanase

Presentation of the manuscript

Many species of filamentous fungi secrete hydrolases that have been successfully applied in the food, feed, textile, pulp and paper industries. *Talaromyces versatilis* (basionym, *Penicillium funiculosum*) has the ability to secrete a very wide range of enzymes capable of hydrolysing the complex polymers found in plant cell walls. This cocktail of enzymes, produced by fermentation, is commercialized by Adisseo under the name of Rovabio™ for use as animal feed additive, increasing the nutritional value of cereal- and soya-based diets (patent WO 99/57325). Although the Rovabio™ cocktail has received permanent authorizations for worldwide use in the diets of several animal species since 2004, Adisseo has engaged in and managed to sustain R&D business seeking to increase not only its awareness, but also its performances.

Talaromyces versatilis has an extremely important potential for the production of enzymes, which can find other applications besides the animal feed. The key to improve the utilization of this microorganism is to have a better understanding of mechanisms that underlie the expression of the genes coding for the glycoside hydrolases and other proteins that compose the cocktail. In this sense, the main objectives of this thesis were:

- To perform an analysis of the genomic sequence of *T. versatilis* in order to catalogue the genes coding for glycoside hydrolases, proteins involved in the secretion pathway and the transcriptional regulators that might be involved in the expression of the first two groups of genes
- To analyse the transcriptome of *T. versatilis* when the mycelia are exposed to different carbon sources and in different genetic backgrounds to characterize the regulatory networks involved in the expression of the glycoside hydrolase-encoding genes
- To optimise the molecular toolbox for the genetic engineering of *T. versatilis* in order to construct deletion mutants for the study of the gene function

The works done during this thesis are presented in the following pages. The manuscript is divided in four main parts: Introduction (chapter I), Materials and Methods (chapter II), Results and Discussion (chapters III – VI) , and Conclusions and perspectives (chapter VII).

The first chapter of this thesis presents different aspects of the fungal biology and also the organism that is object of this work, *Talaromyces versatilis*, and the enzymatic cocktail product of its fermentation, Rovabio™. In the first section, a bibliographic study on the fungal biology, fungal development and taxonomy and the characteristics of *T. versatilis* is presented. A summary of the fungal genomes sequenced lately is done and the *T. versatilis* genome is presented. The plant cell wall composition and its enzymatic deconstruction are introduced in the second section of this chapter. The industrial relevance of the hydrolytic enzymes and in particular those produced by *T. versatilis* are discussed as well. The known regulatory mechanisms that control the expression of the hydrolytic enzyme encoding genes in fungi are presented in the following section.

In the last section an introduction on the tools for the study of the genomics and transcriptomics in fungi is done. The advances in sequencing technologies are presented, in particular of those technologies that have been applied in the study of the genome or the transcriptome of *T. versatilis*. Then, the reference genes for RT-qPCR data normalisation are introduced. Finally, an overview on the existing systems for the genetic modification of filamentous fungi is presented.

The second chapter presents the materials and methods used throughout this work, including the mycelia cultures and media, the molecular biology techniques, the different tools to study the transcriptome and the statistical analysis tools.

The four chapters that follow correspond to the results and discussion. In chapter 3 a brief study of the *T. versatilis* industrial strain growth and development was done, and it was compared to other filamentous fungi: *Trichoderma reesei*, *Aspergillus niger* and *Aspergillus nidulans*. In this section, the

marked capacity of *T. versatilis* to use complex carbon sources, including lignocellulosic material, was shown.

In chapter 4, the analysis of the genomic annotation is introduced. This chapter deals with the manual curation of three main groups of genes, namely, glycoside hydrolases encoding genes, transcription regulators encoding genes and genes involved in the secretion pathway. The difficulties presented by the bad quality of the automatic annotation and the use of the transcriptomic results for the improvement of that annotation are discussed herein.

The chapter 5 deals with the transcriptomic analysis of *T. versatilis*. In the first part, the RNA-seq genome wide study is presented, focusing in the groups of genes of interest: glycoside hydrolase encoding genes, transcriptional regulator encoding genes and genes involved in the secretion pathway. Due to the bad quality of the automatic annotation, the mapping of the RNA-seq results was difficult and a second mapping was done over the improved annotation, these results are also shown.

The second part of this chapter consists in the identification and validation of the reference genes for RT-qPCR data normalisation. The validation of the reference genes was extended beyond *T. versatilis*, to other phylogenetically distant fungi, by using publicly available transcriptomic data. This work issued a publication that was submitted to BMC Genomics which is presented in its entirety in the Appendix II. And in the third part of chapter five the RT-qPCR study of a few GOIs in the industrial strain exposed to a variety of culture conditions and the classification of the GOIs according to their expression patterns are presented.

Chapter 6 introduces the deletion system adapted to *T. versatilis*. This work also issued a publication in Applied and Environmental Microbiology, in collaboration with Professor Archer' laboratory. The deletion system allowed the production of two *T. versatilis* mutants. The expression

of the GOIs by RT-qPCR was done for the $\Delta x/nR$ strain, and the results are discussed in the second part of this chapter.

The seventh chapter is a resume of the main results and conclusions and the perspectives that can be developed in the future to continue this line of investigation.

1. General introduction

1.1. Fungal biology

1.1.1. Fungal growth and development

1.1.1.1. Spore formation and dispersal

The spores or conidia are the unit of propagation for most filamentous fungi. The spores develop, grow and expand over a substrate and form the mycelium composed of hyphae. A fungal spore can be defined in general as “microscopic propagules that lack an embryo and are specialized for dispersal or dormant survival” (Deacon, 2006). Spores can be produced by either a sexual cycle or an asexual cycle. Both sexual and asexual spores function in dormant survival and serve for spreading of the organism. Since *T. versatilis* has no sexual cycle known so far (Figure 1.7), and only the asexual conidia were used throughout this work, only the asexual spores will be discussed in this section.

The spores have some common characteristics: They have low respiration rate, low rates of protein and nucleic acid synthesis. The spore wall is often thicker, with additional layers or additional pigments, the cytoplasm is dense and they have relatively low water content. Finally, spores have a high content of energy storage materials such as lipids, glycogen, or trehalose (Deacon, 2006). The pigments usually present protect the spores from desiccation, osmotic lysis and UV radiation.

In filamentous fungi, the formation of the asexual spores or conidia is triggered by several environmental factors. The nutritional status of the fungus can be a signal; carbon or nitrogen starvation leads to the conidiophore development in *Aspergilli*, but light is another important factor and a pheromone system is also involved (Reinhard Fischer, 2002). In *Trichoderma* species, conidiation is triggered by UV or blue light, endogenous circadian rhythms, type of carbon and

nitrogen source present and the C:N ratio, low ambient pH, the presence of extracellular calcium or the physical injury to the hyphae (Steyaert *et al.*, 2010).

Conidia can be produced by thallic or blastic development. The first one consists essentially of a fragmentation process involving septation. Whereas the blastic type conidia are formed by a budding or swelling process and then the spore become separated from the parent cell. Most conidia are blastic type. The process starts by the formation of a conidiophore from a specialised foot cell hyphae. The conidiophore produces flask-shaped phialides each of which is uninucleate. The phialide extrudes a spore from its tip, and during this process the nucleus divides so that one daughter nucleus enters the developing spore while the other nucleus remains in the phialide to repeat this process. This results in a chain of conidia at the tip of the phialide cell. The process is similar in *Aspergillus*, *Penicillium* or even *Trichoderma* species, but there are variations in the ways that conidiophores and phialides are arranged. *Penicillium*, for instance, present brush-like conidiophores, terminating in phialides, and the successive conidia accumulate in chains (Figure 1.1) (Deacon, 2006). *Talaromyces* species present a very similar conidiophore and phialides shape to *Penicillium* species.

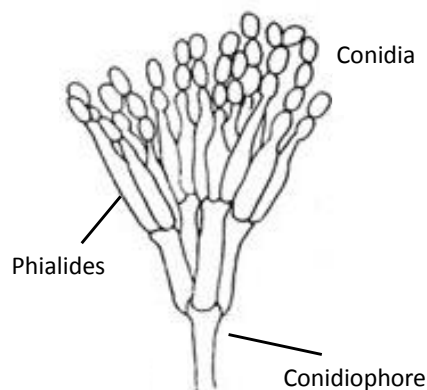


Figure 1.1: Representation of the conidiophore and asexual conidia in *Penicillium* sp.

Conidia can be easily dispersed by wind or rain-splash. Once the spores are in the air, their final destination depends on the meteorological factors, such as wind speeds and rain. The long-distance

dispersal is determined by the resistance to desiccation and the resistance to ultraviolet radiation, conferred by wall pigments. Finally, spores can be deposited by three different methods, namely, sedimentation, impact or washout (Deacon, 2006).

1.1.1.2. Spore germination and vegetative growth

Asexual spores remain dormant if the environment is unsuitable for growth, this is called exogenously imposed dormancy, but they will germinate readily in response to the presence of nutrients such as glucose. In laboratory conditions, most asexual spores germinate promptly at suitable temperature, moisture, pH and oxygen levels. When triggered to germinate, all spores follow the same process. In the first phase of germination, dormancy is broken by environmental cues such as the presence of water and air either or not in combination with inorganic salts, amino acids or fermentable sugars. Then, in the second phase, the cell swells and becomes hydrated, there is a marked increase in respiratory activity, followed by a progressive increase in the rate of protein and nucleic acid synthesis. During this stage, the diameter of the spore increases two fold or more due to water uptake and the viscosity of the cytoplasm decreases. A germ tube is formed by polarised growth in the third phase. And later, the growth speed of the germ tube increases and it eventually develop into a hypha (Deacon, 2006; Krijgheld *et al.*, 2013).

A colony can result from a single sexual or asexual spore that germinates as described before. The process of nuclear replication and segregation into a new septated compartment is defined as duplication cycle. However, the apical compartment of the hyphae is often multinucleate. Hyphae commonly branch at some distance behind the leading growing tip. The branching process is not fully understood yet, but the branching spot is defined by the appearance of a Spitzenkörper at the site of tip emergence. The branching allows the fungus to fill the space efficiently, especially in a nutrient-rich substrate, where branching is more frequent in order to exploit the resources efficiently. When the nutrients are not enough, the branching frequency is lower, producing an effuse mycelium for resources exploration (Kavanagh, 2011). Colonies can reach a diameter in the millimetre scale, which

are called micro-colonies, or in the centimetre scale, and they are called macro-colonies. This depends on the size and composition of the substrate. Aspergilli, as well as *T. versatilis*, produce radial symmetrical macro-colonies in agar plates (Krijgsheld *et al.*, 2013).

1.1.1.3. *Hyphae and apex structure*

A hypha consists of a rigid cell wall that contains the protoplasm, its length is indeterminate, but the diameter is fairly constant, from 2 to 30 μm , depending on the species and the growth conditions. The cell wall is a dynamic exoskeleton that protects the protoplast and defines the growth direction, cell strength and shape (Figure 1.2). The cell membrane and wall are in intimate contact, leaving no periplasmic space, this makes the filamentous fungi resistant to plasmolysis. The fungal membrane contains ergosterol as the main sterol as opposed to cholesterol in animals and phytosterol in plants. Most fungi have cross wall, called septa, at regular intervals. The septa have pores through which the cytoplasm can migrate, and even the nuclei pass through those pores; hence the hyphae can be defined as a chain of interconnected compartments. The wall is thinner in the apex and thicker behind the tip, in the older zones of the hyphae (Deacon, 2006; Kavanagh, 2011).

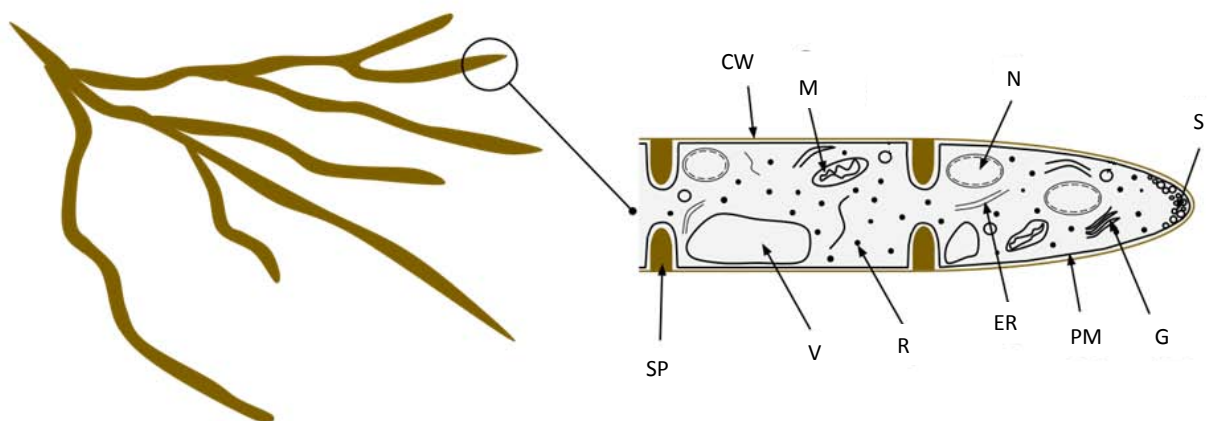


Figure 1.2: Schematic representation of a fungal hypha showing the main structures of the fungal cell. CW: cell wall; M: mitochondria; N: nucleus; S: Spitzenkörper; G: Golgi

apparatus; PM: plasma membrane; ER: endoplasmic reticulum; R: ribosome; V: vacuole;
SP: septum.

Fungal cell wall consists mainly of polysaccharides, proteins and other components are present in a lower proportion. The polysaccharide components can be divided into two types: the structural or fibrillar polymers and the matrix components (Deacon, 2006). Fibrillar or structural polymers are straight chain molecules and consist mainly of chitin cross-linked to glucan molecules in ascomycetous fungi and the matrix component cross-links the fibrils and surrounds the structural polymers. Chitin is composed of long chains of β -1,4 linked N-acetylglucosamine residues, whereas glucans are polymers of glucose joined by β -1,3- or β -1,6-linkages. The chitin chains are mainly located in the inner region of the cell wall and overlaid by the matrix components which cross-link and further strengthens the structure (Deacon, 2006; Latgé & Beauvais, 2014; Munro, 2013).

The septa, which appear at regular intervals, have a pore that allows the traffic of protoplasm along the hyphae. However, if a hyphae is damaged, a Woronin body can block the pore and seal the hyphae leaving the damaged region localised to one area. Most importantly, septa are associated with the differentiation process; by blocking the septal pores, a series of continuous compartments can become independent cells or regions that can undergo separate development (Deacon, 2006).

Hyphae clearly show a polarised growth that occurs mainly at their tips, where there is a region named the extension zone that can be up to 30 μ m long in the fast-growing hyphae such as *Neurospora crassa*. Behind the growing tip, the hypha ages progressively and in the oldest regions it may break down by autolysis or be broken down by the enzymes of other organisms (heterolysis). While the tip is growing, the protoplasm moves continuously from the older regions of the hypha towards the tip. At the growing tip, there is a large accumulation of membrane-bound vesicles with different contents (Figure 1.3). The vesicles are probably originated at the Golgi apparatus, which is found further back in the apical region. This assembly of vesicles in the tip of the hyphae is called apical vesicle cluster (AVC) or Spitzenkörper. Ribosomes were also found in the Spitzenkörper,

suggesting that translation and protein synthesis also happens in the apical region. Even though the Spitzenkörper is composed mainly of vesicles, its centre is vesicle-free and consists of a network of actin filaments. Right behind the tip there is a zone with few or no major organelles, but that is rich in mitochondria, which provide the proton motive force. Septate fungi, in general, have several nuclei in the apical compartment, but sometimes only one or two nuclei in compartments behind the apex, though the nuclei can migrate through the pore of the septa (Deacon, 2006; Steinberg, 2007).

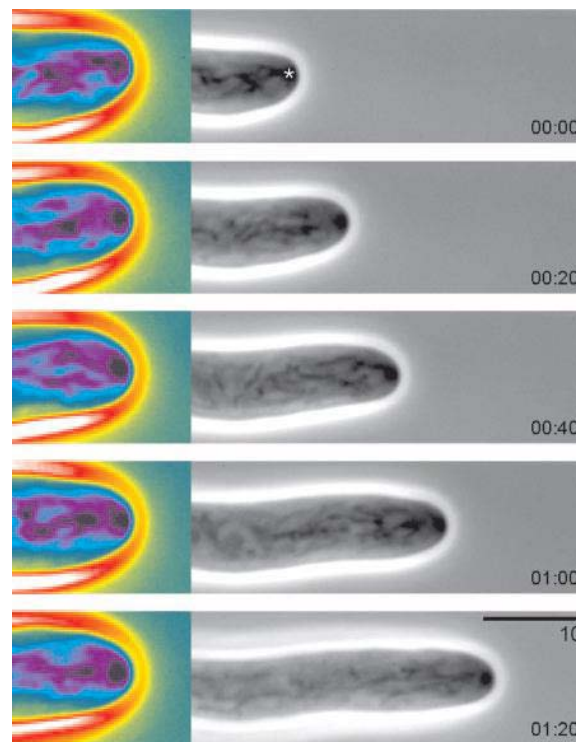


Figure 1.3: Tip growth and the Spitzenkörper in *Neurospora crassa*. Rapid hyphal elongation through 1% low-melting agarose is accompanied by a dynamic accumulation of vesicles in the apex (blue in pseudo-colored image on the left, dark in phase-contrast micrographs on the right). Elapsed time is given in minutes:seconds. Bar, 10 μm . Taken from Steinberg (2007).

1.1.1.4. The secretion pathway

Fungi are exodigesters, which means that they need to produce extracellular enzymes that can efficiently degrade the biopolymers present in their host or in their environment, so then the fungus can incorporate the small molecules product of the degradation. Saprophytic fungi, such as *T.*

versatilis, produce enzymes for the degradation of the polymers in the environment to obtain their nutrients (Girard *et al.*, 2013).

The fungal secretory system is composed of the endoplasmic reticulum (ER), the Golgi apparatus or Golgi equivalent, and the membrane-bound vesicles. The ER consists of interconnected sacs and tubules that form the ER lumen. The entrance of a protein to the secretion pathway is determined by the presence of a signal peptide that targets the protein to the ER where the process begins. The proteins that contain that signal sequence are imported into the ER lumen either co-translationally, *i.e.* the translation occurs simultaneously with the import into the ER through an ER associated ribosomes, or post-translationally, *i.e.* the translation occurs first and the unfolded protein is imported into the ER once the protein has been released from the cytosolic ribosome. The co-translational translocation depends on the presence of a signal recognition particle (SRP) that recognizes and binds the protein at the N-terminus, whereas the post-translational translocation is independent of the SRP (Archer & Turner, 2006). For both kinds of translocation the transport into the ER lumen happens via a translocator complex located in the ER membrane. The complex is formed by several proteins, including SEC61 that form a channel through the membrane. The SPR that binds the peptide being synthesised by the ribosome directs the mRNA-ribosome-peptide complex to the translocation channel in the ER membrane, the translocation occurs, and the translation continues with the peptide translocated into the ER lumen (Figure 1.4) (Saloheimo & Pakula, 2012).

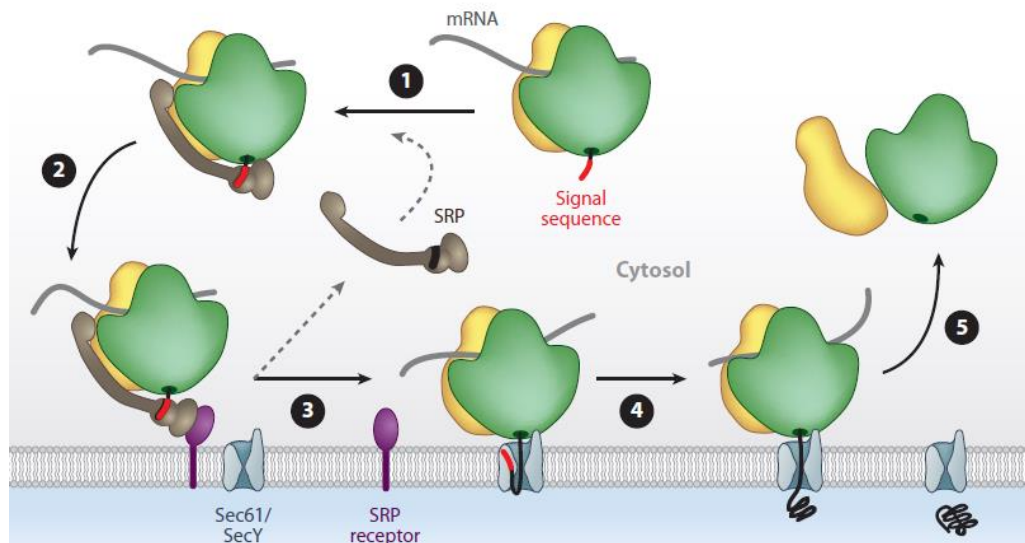


Figure 1.4: Model of co-translational translocation. Step 1: Binding of the signal recognition particle (SRP) to a ribosome carrying a nascent polypeptide with an exposed signal sequence. Step 2: Binding of the mRNA-ribosome-peptide –SRP complex to the SRP receptor. Step 3: Release of SRP, binding of the ribosome to the Sec61 channel, and transfer of the peptide into the channel. Step 4: Translocation of the polypeptide chain, signal sequence cleavage, and folding of the polypeptide on the other side of the membrane. Step 5: Termination of translocation and dissociation of the ribosome. Taken from Park & Rapoport (2012).

For the post-translational translocation in yeast, the unfolded peptide associated with cytosolic chaperones interacts with the SEC61-SEC62-SEC63 complex in the ER membrane and the cytosolic chaperones are released. Once the polypeptide is inserted into the SEC61 channel, its translocation occurs and it binds to the BiP chaperone inside the lumen of the ER that helps the translocation to take place (Figure 1.5) (Park & Rapoport, 2012). The BiP chaperone binds to unfolded protein and holds the protein in an unaggregated state that affords a helpful folding environment. The SIL1 protein reduces the affinity of BiP for the newly synthesised protein and provokes its release by stimulating the ADP:ATP exchange of BiP. LHS1 and KAR2 (or BiP) are Hsp70 chaperones that reside into the ER and have been identified in *S. cerevisiae* and orthologues are found in other fungi, such as the Aspergilli. Other proteins are also involved in the correct folding of the proteins that enter the ER, like the PDI, the ERV1 family, the EUC and the EPS, for instance. Newly synthesised polypeptide

chains are not only correctly folded within the ER but are also glycosylated at the same time by the N- or O-glycosylation of asparagine or serine and threonine residues. Sugar nucleotides are transported from the cytosol into the ER where enzymes such as glucosidases I and II as well as several ER-mannosidases add and cleave glycosyl residues (Geysens *et al.*, 2009).

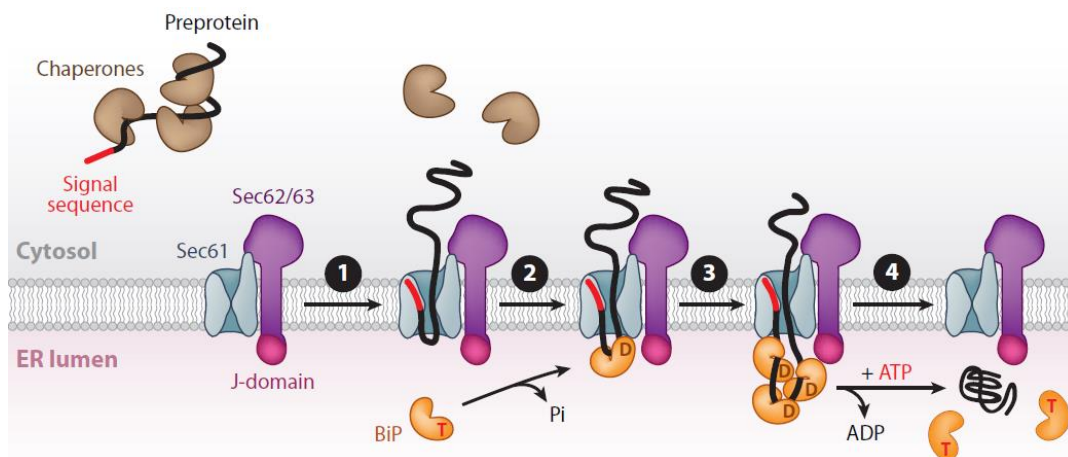


Figure 1.5: Model of post-translational translocation in eukaryotes. Step 1: Binding of a completed polypeptide chain to the SEC complex, consisting of the SEC61 channel and the SEC62/SEC63 complex. Release of the chaperones associated with the polypeptide. Step 2: BiP interacts with SEC63 and then binds to the polypeptide, preventing the polypeptide from sliding back into the cytosol. Step 3: When the polypeptide chain has moved a sufficient distance into the ER lumen, the next BiP molecule binds. Step 4: Nucleotide exchange releases BiP from the polypeptide chain. Taken from Park & Rapoport (2012).

Correctly folded proteins are targeted to a COPII vesicle that incorporates them and moves them to the *cis*-Golgi. Further glycosylation events, catalysed by other α -mannosidases occur within the Golgi apparatus, which in fungi consists of sausage shaped-strings, beads and loops, and hence has a different structure from the animal or plant Golgi. Once the proteins have passed through the Golgi and have arrived to the *trans*-Golgi, they can be integrated into vesicles that will transport them to the membrane. The vesicles fuse with the membrane mainly in the apex and release their cargo. All the vesicle transport, including the ER to Golgi and the Golgi to membrane vesicle transport, occurs in a cytoskeleton-assisted process with vesicles coated by specific proteins (Archer & Turner, 2006;

Spang, 2008). A representation of the secretion pathway with the different steps indicated is shown in Figure 1.6.

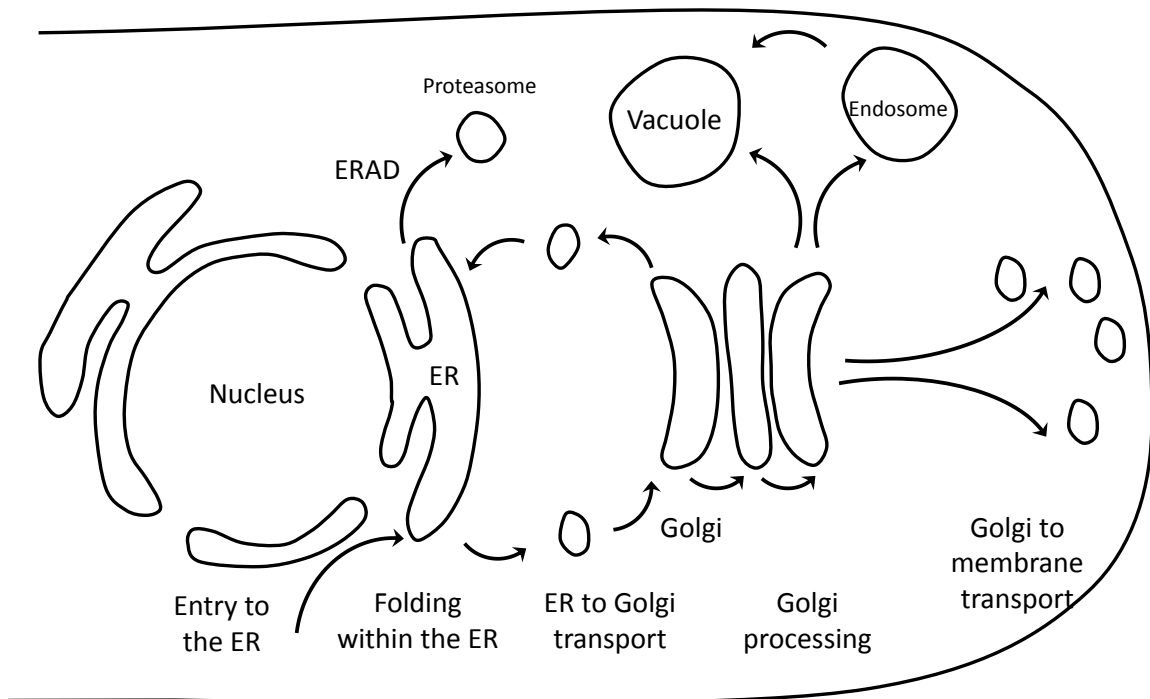


Figure 1.6: Scheme of the secretion pathway in filamentous fungi. ER; endoplasmic reticulum. ERAD; ER associated degradation.

The proteins have to pass rigorous quality controls before continuing their way to the membrane. The ER has quality control systems that monitors the folding state of the proteins and sends the misfolded or aggregated proteins to be degraded (Saloheimo & Pakula, 2012). The presence of misfolded proteins induces the unfolded protein response (UPR) that mediates the up-regulation of ER chaperones, foldases and the ER-associated degradation system (ERAD). The unfolded proteins are retro-translocated to the cytoplasm, ubiquitinated and targeted for degradation by the proteasome. The UPR is mediated by the Hac1/A transcription factor and it is presented later on, in section 1.3.

1.1.2. *Talaromyces versatilis* in the fungal classification

The phylogeny of the *Trichocomaceae* lineage was recently studied and based on a four-gene phylogeny, where 3 different families were defined, namely, *Aspergillaceae*, *Thermoascaceae* and *Trichocomaceae* (Houbraken & Samson, 2011). In fact, the taxonomy of both the *Penicillium* and *Talaromyces* genera was also revised by Samson *et al.* (2011). The *Talaromyces* genus is classified in the *Ascomycota* division, class *Eurotiomycetes*, order *Eurotiales*, and family *Trichocomaceae*. It is composed of species that show yellow or red colonies caused by the accumulation of azaphilones or anthraquinones characteristic to this genus (Osmanova *et al.*, 2010). They have septate interwoven hyphae. They present conidiogenous cells as phialidic, aculeate or acerose, from which conidia, generally green, appear in basipetal connected chains, usually ellipsoidal to fusiform. The species that present a sexual cycle have ascomatal cleistothesia, which present asci with 8 spores each, globose to ellipsoidal. The mature asci are produced in chains. The *Talaromyces* genus, as well as the *Penicillium* and *Aspergillus* genera, is composed of industrially important fungi, although the *Talaromyces* species are less frequently used (Houbraken *et al.*, 2014).

The fungus studied in this work was first classified within the *Penicillium* genera. But recently, the molecular characterization defined it as a *Talaromyces*. Hence, *Penicillium funiculosum* adopted the *Talaromyces versatilis* name of species (Bridge & Buddie, 2013).

T. versatilis is a mesophilic, filamentous fungus that can grow aerobically in a diversity of organic substrates. It can be isolated from a number of environments. The ability to produce a wide variety of enzymes allows the development over a diversity of organic substrates. This fungus does not have a sexual cycle known so far. In fact, the asexual production of conidia is the only dispersion way currently known (Figure 1.7). Conidia can be dispersed by the wind, water or insects and if they find a propitious environment, they can germinate to develop a mycelium and start the cycle once again.

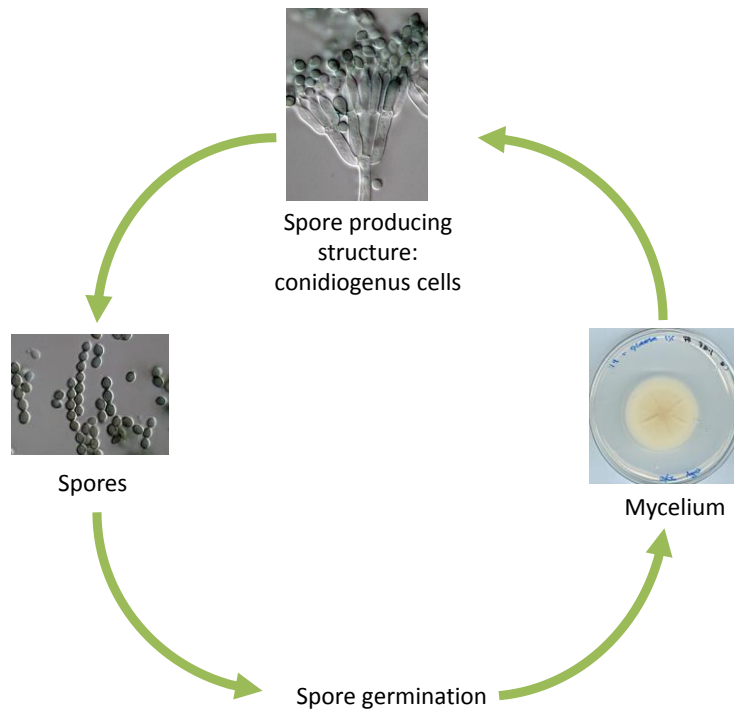


Figure 1.7: Representation of the asexual life cycle of *Talaromyces*. The conidia and conidiogenous cells pictures were taken from Frisvad *et al.* (2013).

1.1.3. Filamentous fungi genomes

Since the development and improvement of the sequencing technologies, the cost of a *de-novo* sequencing of a genome has decreased importantly. Therefore, the sequencing of the genomes of several filamentous fungi became possible. The genomes of yeast species, such as *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996), were first sequenced. But the sequencing of filamentous fungi readily started soon with model fungi, such as *Neurospora crassa* (Mannhaupt *et al.*, 2003), *Aspergillus nidulans* (Galagan *et al.*, 2005), or industrially important fungi, like *Trichoderma reesei* (Martinez *et al.*, 2008; Seidl *et al.*, 2008), *Penicillium chrysogenum* (van den Berg *et al.*, 2008), or even pathogens, such as *Aspergillus fumigatus* (Nierman *et al.*, 2005). In total, over a hundred fungal genomes have been sequenced so far. This number seems outrageously small if we take into account that the latest estimation of the total number of fungal species in the world is of about 3.5 to 5.1 million, most of which are still unknown (Blackwell, 2011).

The already known fungal genomes show a very wide diversity in their sizes, number of chromosomes, content of transposable elements (Stukenbrock & Croll, 2014). The fungal genomes listed in Table 1.1 correspond to some examples of the genomes sequenced lately amongst the Ascomycetes and Basidiomycetes. The variety in the genome size and protein coding gene number is remarkable. The genome size varies from 27 Mb to over 100 Mb, and the number of predicted genes varies from 5,000 to 23,000. But the size of the genome is not always related to the number of genes, for instance, the *B. graminis* genome is one of the biggest in the list, with 120 Mb, but has the lowest amount of protein coding genes (5,854). The GC content in most of the fungi in the list is around 50%, although *R. irregularis* that has a very low GC content (33%) and *Phanerochaete chrysosporium* that has a high GC content (57%) constitute the exceptions. The diversity observed here is a reflection of the great diversity within the fungi kingdom.

Table 1.1: List of recently sequenced fungal genomes with their metrics.

Species	Genome size	GC content	Protein coding genes	Reference
<i>Neurospora crassa</i>	38.7 Mb	49.6%	10,620	(Galagan <i>et al.</i> , 2003)
<i>Magnaporthe grisea</i>	37.9 Mb	51.6%	11,109	(Dean <i>et al.</i> , 2005)
<i>Aspergillus nidulans</i>	30.1 Mb	50%	9,541	(Galagan <i>et al.</i> , 2005)
<i>Aspergillus fumigatus</i>	27.9 Mb	49%	9,926	(Galagan <i>et al.</i> , 2005)
<i>Aspergillus oryzae</i>	37 Mb	48%	14,063	(Galagan <i>et al.</i> , 2005)
<i>Aspergillus niger</i> CBS 513.88	34 Mb	50.4%	14,082	(Andersen <i>et al.</i> , 2011; Pel <i>et al.</i> , 2007)
<i>Fusarium graminearum</i>	36.1 Mb	48.3%	11,640	(Cuomo <i>et al.</i> , 2007)
<i>Trichoderma reesei</i>	33.9 Mb	52%	9,129	(Martinez <i>et al.</i> , 2008)
<i>Phanerochaete chrysosporium</i>	29.9 Mb	57%	11,777	(Martinez <i>et al.</i> , 2004)
<i>Penicillium chrysogenum</i>	32.2 Mb	48.9%	12,943	(van den Berg <i>et al.</i> , 2008)
<i>Blumeria graminis</i> sp. <i>hordei</i>	120 Mb	44%	5,854	(Spanu <i>et al.</i> , 2010)
<i>Sordaria macrospora</i>	39.8 Mb	52.4%	10,789	(Nowrousian <i>et al.</i> , 2010)
<i>Aspergillus niger</i> ATCC1015	34.9 Mb	50.4%	11,200	(Andersen <i>et al.</i> , 2011)
<i>Leptosphaeria maculans</i>	45.1 Mb	44.1%	12,469	(Rouxel <i>et al.</i> , 2011)

<i>Colletotrichum graminicola</i>	50.9 Mb	49.1%	12,006	(O'Connell <i>et al.</i> , 2012)
<i>Colletotrichum higginsianum</i>	49.1 Mb	55.1%	16,172	(O'Connell <i>et al.</i> , 2012)
<i>Fibroporia radiculosa</i>	30.9 Mb	53.8%	9,262	(Tang <i>et al.</i> , 2013)
<i>Ganoderma lucidum</i>	43.3 Mb	55.9%	16,113	(Chen <i>et al.</i> , 2012)
<i>Rhizophagus irregularis</i>	101 Mb	33%	23,561	(Tisserant <i>et al.</i> , 2013)
<i>Pyronema omphalodes</i>	50 Mb	47.8%	13,369	(Traeger <i>et al.</i> , 2013)
<i>Fusarium fujikuroi</i>	43.9 Mb	47.4%	14,813	(Wiemann <i>et al.</i> , 2013)
<i>Fusarium verticillioides</i>	41.8 Mb	48.6%	14,180	(Wiemann <i>et al.</i> , 2013)
<i>Fusarium oxysporum</i>	61.4 Mb	47.3%	17,458	(Wiemann <i>et al.</i> , 2013)
<i>Fusarium circinatum</i>	44.3 Mb	47.3%	15,022	(Wiemann <i>et al.</i> , 2013)
<i>Fusarium mangiferae</i>	45.6 Mb	48.8%	16,261	(Wiemann <i>et al.</i> , 2013)
<i>Fusarium solani</i>	51.3 Mb	50.7%	15,702	(Wiemann <i>et al.</i> , 2013)
<i>Pyrenochaeta lycopersici</i>	54.8 Mb	39.4%	17,411	(Aragona <i>et al.</i> , 2014)
<i>Rhizoctonia solani</i> AG8	39.8 Mb	50%	13,964	(Hane <i>et al.</i> , 2014)
<i>Rhizomucor miehei</i>	27.6 Mb	43.8%	10,345	(Zhou <i>et al.</i> , 2014)

Currently, large scale community efforts are being done to carry on projects like the 1000 Fungal Genomes Project from the JGI (<http://1000.fungalgenomes.org>) or the Fungal Genome Initiative from the Broad Institute (<http://www.broadinstitute.org/scientificcommunity/science/projects/fungal-genome-initiative>). This kind of projects will provide a larger number of fungal genomes to be studied and compared (Stukenbrock & Croll, 2014).

The 20 known Aspergilli genomes, whose sequences are grouped in the AspGD database (<http://www.aspgd.org/>), have enough evolutionary distance visible in the divergence of their sequences, but are close enough so that the orthologs and syntenic regions can be identified. The availability of a multi-species database has allowed to expand the curation of the genomes in an effort to maintain and update the gene annotation of all these species (Cerqueira *et al.*, 2014). The AspGD database contains the full genomes of *Aspergillus fumigatus* A1163, *Aspergillus fumigatus*

Af293, *Aspergillus nidulans* FGSC A4, *Aspergillus niger* CBS 513.88 and *Aspergillus oryzae* RIB40. For the other species, only the coding sequences are available.

1.1.4. *Talaromyces versatilis* genome

In an attempt to have a better understanding of the biology of *T. versatilis* and to evaluate its potential as an enzymes producing fungus, Adisseo decided to sequence its genome. To start with, a first sequencing was performed using the Roche/454 pyrosequencing technology. Over 800,000 individual lectures were obtained from this sequencing that could be assembled in 36,581 contigs. But only 12,524 contigs, with an average length of 2 kb, were kept and used later for the annotation of the genome (Guais, 2009). Later on, Adisseo decided to upgrade the poor quality of the first sequencing by doing a second one, this time with the Solexa/Illumina technology. With this second sequencing, the genome of *T. versatilis* was improved greatly; 751 contigs could be reconstructed, with an average length of 46.8 kb. The total sequences add up to 35.16 Mb, with a GC content of 46%. An automatic annotation was done over the new sequence giving 18,670 genes in total, from which there are 141 genes that code for a tRNA gene and only one that codes for an rRNA gene 18S. This indicated that, even though the sequence was improved significantly, it is still incomplete and/or not correctly annotated. The genome statistics obtained in each of the sequencings are detailed in Table 1.2.

Table 1.2: Genome statistics of *T. versatilis*.

	First sequencing – Roche 454	Second sequencing – Illumina
Contigs	12,524	751
Average contig length	2 kb	46.8 kb
Base-paired sequenced	36 Mb	35.16 Mb
GC%		46%
Protein encoding genes		18,670
tRNA encoding genes		141
rRNA encoding genes		1 (18S)

1.2. Plant polymers degradation by filamentous fungi

1.2.1. Industrial application of fungal enzymes

Fungi are used in different industries, including food, beverage and feed industries. The fungal products applied in these industries, which can be enzymes or metabolites, are either “additives” or “processing aids” (Archer *et al.*, 2008). The additives are provided with a particular purpose and remains in the food, although it is not a characteristic ingredient. Whereas the processing aids, which are mainly the enzymes, are also provided for a determined reason and remain in the food, but only as residues.

The fungi generally used for the production of enzymes are capable of degrading polymers on their own. They are able to produce enzymes for the degradation of proteins, cellulose, hemicellulose, starch or even lignin, which are applied in detergents, dairy, animal feed, starch processing and textile industries, for instance. The application of enzymes in the feed industry has been spread in the last years. Supplying the monogastric animal feed with enzymes improves the availability of nutrients, and eases the digestion of the cereals (Filer, 2003). Some *Aspergillus sp.*, *Trichoderma sp.* and *Penicillium sp.* are frequently used in industry to produce polysaccharide degrading enzymes. *Rhizopus sp.* and *Aspergillus sp.* are, on the other hand, important for the production of other enzymes, such as glucoamylase, lipase and pectinase (Archer *et al.*, 2008).

Another important industrial application of fungal enzymes is their implication as pre-treatment of biomass in biofuel production and, more largely, for biorefinery purposes (Himmel *et al.*, 2007). The biofuel production consists of four main steps, namely: 1) pre-treatment of the lignocellulosic material, 2) enzymatic hydrolysis to release sugar residues, 3) sugar fermentation, generally by yeast and 4) distillation and purification of the ethanol to reach the fuel specifications (Margeot *et al.*, 2009). One of the greatest obstacles in this procedure is the hydrolysis of the plant biomass to

fermentable sugars. Fungi are, so far, the only organisms capable of producing the amounts of enzymes needed to complete the enzymatic hydrolysis at an industrial scale (Xu *et al.*, 2009).

1.2.2. Genetic potential of filamentous fungi for lignocellulosic material deconstruction

The hydrolysis of the components of the plant cell wall requires the combined action of multiple enzymes with different activities and substrate specificity. In nature, several microorganisms are capable of producing such enzymes that can catalyse the hydrolysis of the carbohydrates bounds present in plants. The filamentous fungi are probably the most effective degraders, since they produce a large variety of enzymes. The recently sequenced fungal genomes allowed the identification of the hydrolytic enzyme-coding genes and the comparison between the genetic potential of different fungi (Zhao *et al.*, 2013). This study revealed that fungi have a great diversity in the number and variety of hydrolytic enzyme encoding genes and that the repertory of enzymes is related to the lifestyle of each particular fungus and its nutritional strategy.

1.2.2.1. Plant cell wall composition

The plant cell wall is mainly composed of polysaccharides such as cellulose, hemicellulose and the aromatic polymer, lignin. There are also small amounts of pectin and proteins (Aro *et al.*, 2005; Lagaert *et al.*, 2009). The major constituent of the plant cell wall is the cellulose, which consists of D-glucose residues with β -1,4 linkages. This is a linear polymer that forms microfibrils by stacking the chains of glucose one on top of the other, forming a structure that confers great rigidity to the cell wall (Dashtban *et al.*, 2009; de Vries & Visser, 2001).

The hemicellulose is the most heterogeneous and second most abundant component of the lignocellulose. Different types of hemicellulose exist in nature, depending on the source. For instance, the xylans, which are found in cereals and hardwood, consist of a β -1,4 linked D-xylose backbone substituted with side groups such as L-arabinose, D-galactose, acetyl, feruloyl or glucuronic

acid. Therefore, the arabinoxylan, which is part of the crop plants, has a backbone composed of D-xylose and L-arabinose residues linked to the backbone. The arabinose residues may have additional linkages to phenolic compounds (Dodd & Cann, 2009; Lagaert *et al.*, 2009). The galactoglucomannan has a backbone of β -1,4 linked D-mannose and D-glucose residues with D-galactose side groups, and is found in soft and hardwood. The hemicellulose molecules interact with the cellulose microfibrils (de Vries & Visser, 2001).

The third main component of the cell wall is the lignin. It is an aromatic alcohol polymer that links the hemicellulose and the cellulose, and constitutes a barrier strengthening the plant cell wall and conferring rigidity to it (Dashtban *et al.*, 2009). Finally, the pectin is an heteropolysaccharide with a backbone of α -1,4 linked D-galacturonic acid residues, and it has 2 different regions: the “smooth” and the “hairy” regions. In the hairy regions the backbone is interrupted by α -1,2 linked L-rhamnose residues, and chains of L-arabinose and D-galactose can be attached to the L-rhamnose residues (de Vries & Visser, 2001).

The plant cell wall is built of microfibrils of cellulose, which is the most abundant component of the cell wall, held in place by the hemicellulose that crosslinks the cellulose fibrils. They are all embedded in the less structured pectins forming a three-dimensional structure, as it is represented in Figure 1.8. There is evidence to suggest that pectins may also be crosslinked to hemicelluloses, phenolic compounds, and to wall proteins. This crosslinking of pectins to other wall components provides additional structural and functional complexity to the wall (Caffall & Mohnen, 2009). The type of hemicellulose and the polysaccharides composition varies from one plant species to another, cell types and differentiation stages.

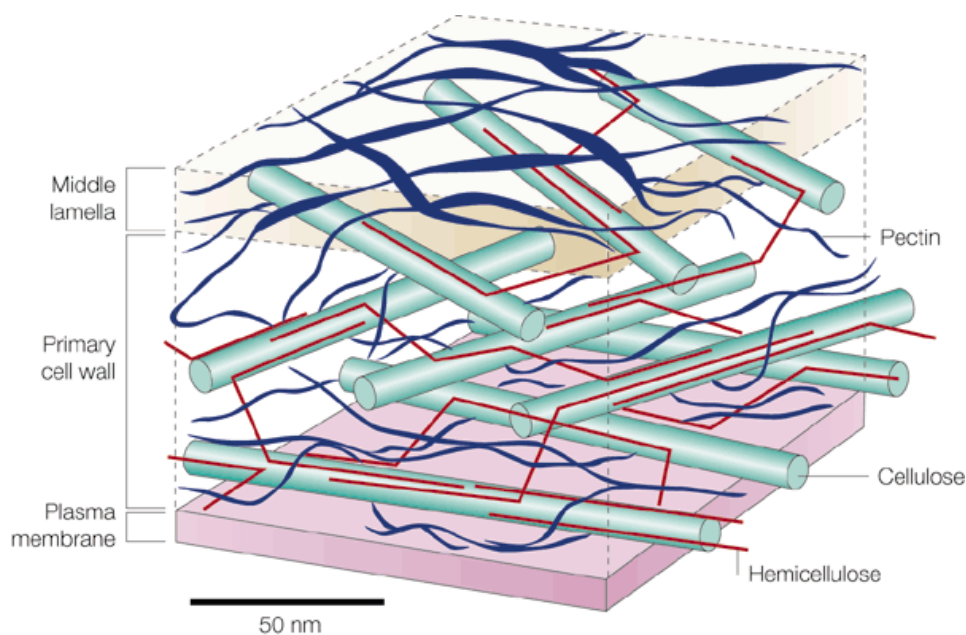


Figure 1.8: Diagrammatic representation of the structure of plant cell wall and associated cellulose, hemicellulose and pectin components. Taken from Smith (2001).

1.2.2.2. CAZymes

The hydrolytic enzymes are part of a larger group of enzymes called carbohydrate active enzymes (CAZymes – www.cazy.org) and are responsible for the assembly, modification and breakdown of the oligo- and polysaccharides (Lombard *et al.*, 2014). They have been classified in five groups, namely: glycoside hydrolases (GH), glycosyltransferases (GT), polysaccharide lyases (PL), carbohydrate esterases (CE) and finally auxiliary activities (AA). The classification was done according to their functional domains and catalytic modules. The GHs, PLs, CEs and AAs are the enzymes involved in the degradation of the different polymers present in the plant cell wall.

The GH group is subdivided in 133 families, based on the amino-acid sequence similarity of the enzymes. Enzymes with different amino-acid sequence, but similar 3D structure and folding can be classified in the same clan. Most of the genes encoding for plant cell wall degrading enzymes in filamentous fungi genomes are classified as GHs. These enzymes hydrolyse the glycosidic bond between two carbohydrates, or a carbohydrate and a non-carbohydrate moiety.

Chapter I

The GHs hydrolyse the glycosidic linkages by two different mechanisms: inversion or retention. Both employ two carboxylic acid residues in the active site and catalyse the reaction by acid-base catalysis. The inversion mechanism consists of a single-displacement reaction where a carboxylic acid group acts as a base to activate a water molecule to attack the anomeric carbon, as a result the anomeric configuration of the substrate is inverted. The retaining mechanism consists in a double displacement involving a first step where one residue acts as a nucleophile, attacking the substrate and form a glycosyl enzyme intermediate and the second residue protonates the glycosidic oxygen. In a second step, the glycosyl enzyme is hydrolysed by water, with the other residue now acting as a base catalyst deprotonating the water molecule as it attacks (Dashtban *et al.*, 2009; Dodd & Cann, 2009).

The GHs are globular proteins that contain a catalytic domain (CD) and some of them have a carbohydrate binding module (CBM). There are 68 CBM families, with modules that have different ligand specificities and several functions. They are responsible for keeping the enzymes near the substrate, and thus accelerate the reaction. The specificity of the CBM makes it possible for the enzyme to be in the vicinity of its target substrate. And finally, some CBMs appear to be able to disrupt the polysaccharides, increasing the degrading capacity of the CD of the protein (Boraston *et al.*, 2004).

PLs and CEs are also involved in the breakdown of glycosidic bonds, but with different mechanisms, and are also classified into families, 23 and 16, respectively, according to their amino-acid sequence. The first ones cleave uronic acid-containing polysaccharides via β -elimination, generating an unsaturated hexenuronic acid and a new reducing end, whereas the CEs catalyse the de-O- or de-N-acetylation of substituted saccharides, releasing sugar residues (www.cazy.org).

The AAs, on the other hand, include the lytic polysaccharide monooxygenases and ligninolytic enzymes. The first group is divided in 3 families, whereas the ligninolytic enzymes are classified in 8 families (www.cazy.org).

In the following sections, the specific enzymes involved in the degradation of the different components of the plant cell wall will be discussed.

1.2.2.2.1. Cellulases

The complete hydrolysis of cellulose to glucose residues requires the action of three different types of enzymes: endoglucanases, cellobiohydrolases and β -glucosidases. The endoglucanases act in the middle of the cellulose chains, and they initiate the cellulose breakdown by attacking the amorphous regions of the cellulose and making it more accessible for the action of other enzymes; indeed, they produce more ends for the cellobiohydrolases to act upon (Dashtban *et al.*, 2009). The cellobiohydrolases can act upon the reducing or the non-reducing end of the cellulose chains end, releasing disaccharides or oligosaccharides, the cellobiose being the main product of these enzymes (Dashtban *et al.*, 2009). The disaccharides and oligosaccharides are the β -glucosidases' substrates, which will converted them to glucose monomers (Aro *et al.*, 2005; Dashtban *et al.*, 2009). The synergistic action of these enzymes is represented on Figure 1.9. The endoglucanases were classified mostly in the GH families 5, 6, 7, 12, 45, the cellobiohydrolases were assigned to the families 6 and 7 and the β -glucosidases to the families 1, 3 and 5.

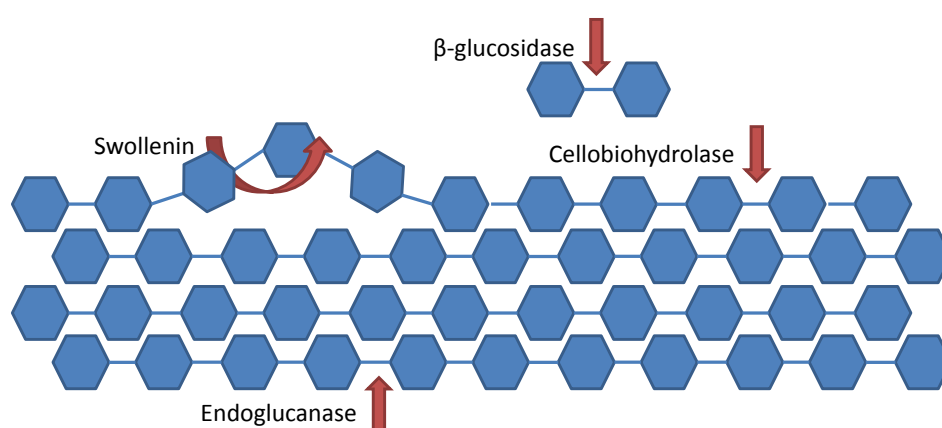


Figure 1.9: Scheme of the cellulose and the 3 types of enzymes that are responsible for its degradation: endoglucanase, cellobiohydrolase and β -glucosidase. The swollenin, which loosens the cellulose fibrils, is also represented.

1.2.2.2.2. Hemicellulases

The heterogeneity that characterises the hemicelluloses demands several types of enzymes for its complete degradation (Dashtban *et al.*, 2009). The hydrolysis of the hemicelluloses is carried out by endo-enzymes that cleave the backbone chain internally, exo-enzymes that release monomers from the ends of the chains and enzymes that cleave the side chains that decorate the backbone. All these enzymes release monosaccharides or oligosaccharides, whose nature depends on the hemicellulose type (Aro *et al.*, 2005).

The β -xylanases cleave the β -1,4 linkage between the xylose residues within the xylan backbone. They are classified in the GH families 5, 7, 8, 10, 11 and 43 based on their amino-acid sequence, folding and catalytic mechanism. The arabinofuranosidases remove the arabinose side chains from the xylan backbone, and they are grouped on GH families 43, 51, 54 and 62. Another enzyme involved in xylan deconstruction is the β -xylosidase, which release xylose residues from the non-reducing end of the xylo-oligosaccharides, which are the product of the xylanases action. They are classified within the GH families 3, 39, 43, 52 and 54. Xylose residues substituted with glucuronic acid are attacked by the α -glucuronidases from the GH family 67. These enzymes release glucuronic acid from the terminal, non-reducing end of the xylo-oligosaccharides. The xylose residues are frequently acetylated or substituted by ferulic acid or coumaric acid residues. The acetylation of the xylose residues affects the capacity for other enzymes, such as the xylanases, to bind and hydrolyse the xylan. The ferulic or coumaric acid esterases cleave the linkages between the acid groups to the backbone, and the acetyl xylan esterases release the acetyl groups. They are classified within the CE family 1 for the ferulic and coumaric acid esterases, and families 1-7, 12 and 16 for the acetyl xylan esterases. The deconstruction of xylans requires the coordinated action of all these enzymes together (Dodd & Cann, 2009) which are represented on Figure 1.10.

The degradation of the galactoglucomannan requires β -endomannanases, which cleave the backbone releasing manno-oligosaccharides and are classified in the CAZy GH family 5. The β -

mannanases produce mannobiose and mannotriose from mannan. Finally, β -mannosidases release mannose residues from the non-reducing end of the manno-oligosaccharides and belong to the GH families 1, 2 and 5 (de Vries & Visser, 2001).

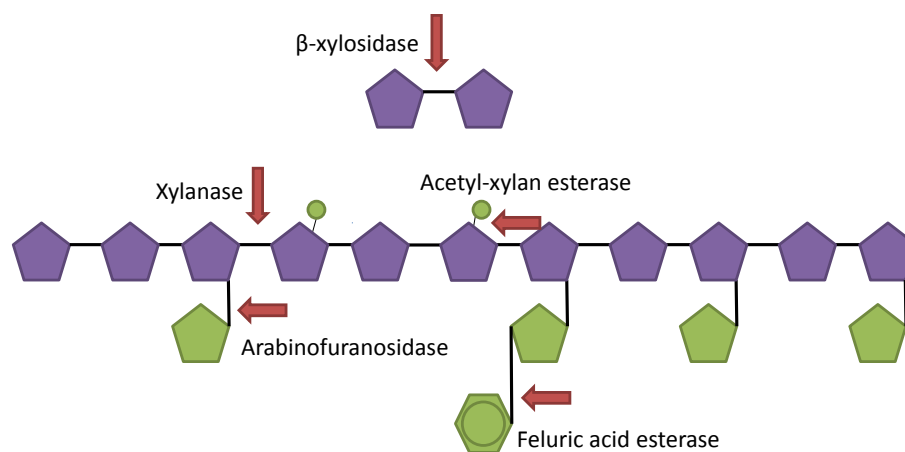


Figure 1.10: Scheme of the arabinoxylan, one of the xylans found in nature, and some examples of the enzymes that are responsible for its degradation: xylanase, β -xylosidase, arabinofuranosidases, ferulic acid esterase, acetyl-xylan esterase.

1.2.2.2.3. *Ligninases*

Lignin is probably the most resilient component of the plant cell wall. Some basidiomycetes seem to be the most effective lignin degraders. However, not all the lignocellulose degrading fungi count with all the necessary enzymes for the degradation of this fraction of the cell wall. The deconstruction of lignin requires, among others, laccases, manganese peroxidases and lignin peroxidases. The laccases oxidize the phenolic compounds of the lignin and reduce the oxygen to water. Both peroxidases catalyse a variety of oxidative reactions that depend on the presence of H_2O_2 (Aro *et al.*, 2005; Dashtban *et al.*, 2009).

1.2.2.2.4. *Auxiliary activities*

The CAZy “auxiliary activities” classification, or AA, includes enzymes that are not necessarily involved in the degradation of carbohydrates, but are necessary for the degradation of lignocellulosic

material (Levasseur *et al.*, 2013). Therefore, this group holds ligninolytic enzymes and also polysaccharide mono-oxygenases.

The former GH61 family members, now part of the AA classification (www.cazy.org), are proteins that, even though they have a very similar structure to that of the true glycoside hydrolases, are not part of this group (Levasseur *et al.*, 2013). The GH61 are copper-dependent oxidoreductases that require redox-active cofactors for their enhanced catalytic activity (Bey *et al.*, 2013). They strongly bind a copper ion, which is essential for their function. These enzymes are unable to fully degrade lignocellulosic material in the absence of other kinds of enzymes (Harris *et al.*, 2010). However, this group of proteins does not need the presence of a free chain of cellulose to act; their strong oxidative power allows them to act over crystalline cellulose microfibrils (Langston *et al.*, 2011; Levasseur *et al.*, 2013). Another characteristic of this group of enzymes is that they are exclusive to lignocellulose degrading filamentous fungi, except for a single bacterial GH61 protein found on *Cryptococcus neoformans* (Harris *et al.*, 2010).

There are other activities outside the CAZy classification, which might be involved in the degradation of lignocellulose, such as swollenins and hydrophobic surface binding proteins. The swollenins have a plant cell wall loosening action, like the plant expansins, and they could render the cellulose microfibrils more available for the cellulases to degrade (Saloheimo *et al.*, 2002). Therefore, the disruptive action of the swollenins can enhance the cellulases activity (Wang *et al.*, 2011). Hydrophobic surface binding proteins are small, surface active proteins with both a hydrophilic and a hydrophobic side. These proteins have different roles in the fungal cell, including structural role, reduction of the surface tension in order to facilitate the interaction with the surfaces, they cover the fungal spores or assemble onto the fungal cell wall (Linder *et al.*, 2005). It has been shown that the hydrophobic surface binding protein encoded by *Aspergillus oyzae* is involved in the interaction with the polybutylene succinate-co-adipate (PBSA), a biologically degradable polymer, and in the interaction with the cutinase, which is responsible for the PBSA degradation (Ohtaki *et al.*, 2006). The

hydrophobic surface binding proteins are potentially involved in the degradation of hydrophobic substrates by facilitating the interaction of the substrate with the degrading enzymes.

1.2.3. Production of hydrolytic enzymes by *Talaromyces versatilis*

The *T. versatilis* industrial strain used for the production of the Rovabio™ cocktail is called *T. versatilis* 8/403, and was originated from the *T. versatilis* IMI123457 strain from the “International Mycological Institute” collection. The industrial strain was obtained by mutagenesis, using UV light exposure, and was selected for being a producer of high amounts of cellulases and hemicellulases. The cocktail naturally produced by *T. versatilis* contains several enzymatic activities, including xylanases, β -glucanases and cellulases, for instance. Previous biochemical studies showed the presence of a variety of enzymes in the cocktail, such as xylanases, esterases, cellulases, β -glucanases, laminarinases, dextranases, arabinanases, mannanases, pectinases and proteases (Guais *et al.*, 2008). What makes this cocktail so effective is probably the combination of enzymes and their synergistic activity to fully degrade the lignocellulosic materials.

1.3. Regulation of the expression of cellulase and hemicellulase genes in filamentous fungi

A few transcription factors controlling the expression of cellulase and hemicellulase encoding genes are known and have been characterized in filamentous fungi. Homologs for most of these transcription factors were found in the genome of *T. versatilis* (see Chapter 4). XlnR, AraR, ClrA, ManR, PacC, ACEII and AreA have been described in different fungi as positive regulators, whereas CreA and ACEI are the main repressors. However, they all belong to the zinc-binding proteins family.

Zinc-binding proteins are one of the largest families of transcriptional regulators distributed amongst the eukaryotes. They have different functions, but they all share the zinc finger motif. There are different classes of zinc-finger proteins, like the C₂H₂ or classic zinc finger, the C₄ zinc finger and the C₆ or Zn(II)₂Cys₆ zinc binuclear cluster proteins. The first two groups are distributed in all the

eukaryotes organisms. Whereas the $Zn(II)_2Cys_6$ binuclear cluster domain (Pfam00172) is a motif widely distributed amongst the known transcriptional activators of the carbohydrate metabolism in *Aspergilli* and are exclusively fungal transcription factors. They possess the motif $CysX_2CysX_6CysX_{5-12}CysX_2CysX_{6-8}Cys$. The two zinc atoms bind to the cysteine in the only zinc finger unit and coordinates the folding and binding to the DNA (MacPherson *et al.*, 2006). They can interact with the DNA as monomers, homodimers or heterodimers (Figure 1.11). As it will be discussed in the following sections, most of the transcription activators known in fungi, and found in *T. versatilis*, belong to this group of proteins.

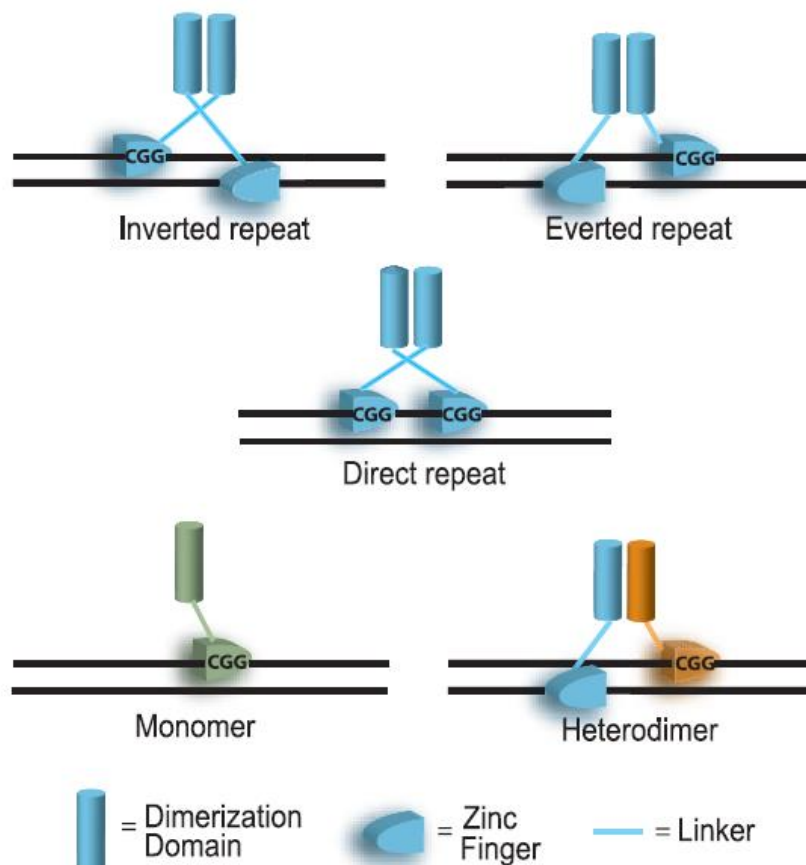


Figure 1.11: Model for the zinc binuclear cluster protein DNA recognition. Zinc cluster proteins preferentially bind to CGG triplets that can be oriented in three different configurations: the inverted, everted, and direct repeats. The orientation of CGG triplets and the nucleotide spacing between the triplets are the two major determinants of DNA binding specificity. The CGG triplet is common, but variations in this motif can be found. Zinc cluster proteins can also bind as monomers (in green) as well as homodimers (two

molecules in blue) and heterodimers (one molecule in blue and one in orange). Taken from MacPherson *et al.* (2006).

1.3.1. The XlnR regulator

The xylanolytic regulator was first described in *A. niger* as required for the regulation of the expression of genes encoding xylanases and xylosidases, but also cellulase genes in the presence of xylan or xylose (van Peij *et al.*, 1998b). Genes coding for *xlnR* homologs have been characterized in other fungi, like *A. oryzae* (Marui *et al.*, 2002), *T. reesei* (Rauscher *et al.*, 2006), *F. graminearum* (Brunner *et al.*, 2007), *F. oxysporium* (Calero-Nieto *et al.*, 2007), *A. nidulans* (Tamayo *et al.*, 2008), *M. oryzae* (Battaglia *et al.*, 2013), *N. crassa* (Sun *et al.*, 2012). The XlnR transcriptional activator is a Zn(II)₂Cys₆ binuclear transcription factor, which is a type of transcription factors characteristic to fungi, described in the previous section. The consensus binding sequence for the XlnR protein has the motif 5' GGCTAA 3', and one to three copies of this motif are present in the promoters controlled by this regulator, as was shown in *A. niger* (van Peij *et al.*, 1998a), *A. oryzae* (Marui *et al.*, 2002), and *T. reesei* (Rauscher *et al.*, 2006).

However, the set of genes controlled by XlnR and the expression patterns of these genes differ from one fungus to the other, for example, in *A. niger*, XlnR controls the expression of endoxylanase, β -xylosidase, endoglucanase and cellobiohydrolase encoding genes (van Peij *et al.*, 1998b, a), whereas in *M. oryzae* it regulates the expression of the genes involved in the Pentose Catabolism Pathway and it does not affect the expression of xylanolytic genes (Battaglia *et al.*, 2013). In *T. reesei* XYR1 is involved, not only in the xylanases and cellulases expression, but also in the metabolism of xylose by its influence on the expression of D-xylose reductase (Stricker *et al.*, 2006). In *F. graminearum*, on the other hand, neither xylose nor xylan was able to activate the xylanase genes, and only the wheat cell walls could induce the up-regulation of these genes. And even though a difference in the transcript levels was detected between the wild-type strain and the *xlnR* mutant, the loss of expression was not as drastic as it can be observed in other fungi. In addition, the

expression of the cellulase genes is not affected by the lack of the XlnR regulator in this plant pathogen (Brunner *et al.*, 2007). In another plant pathogen, *F. oxysporium*, XlnR also controls the main xylanase genes, but in its absence the fungus still can degrade xylan and its virulence is not affected; the possibility that in this case XlnR does not control the expression of all the xylanase genes has to be considered (Calero-Nieto *et al.*, 2007). Conversely, overexpression of *xlnR* leads to increased expression of the genes under its control (Battaglia *et al.*, 2013; Brunner *et al.*, 2007; Gielkens *et al.*, 1999; Marui *et al.*, 2002; Noguchi *et al.*, 2009; Stricker *et al.*, 2006; Sun *et al.*, 2012).

1.3.2. The AraR regulator

Battaglia *et al.* (2011a) identified a regulator closely related to XlnR, but that responds to the presence of arabinose in *A. niger*. This new regulator was called AraR. It also contains a Zn(II)₂Cys₆ binuclear cluster domain (Pfam00172), and a fungal specific transcription factor domain (Pfam04082). The AraR protein has 32% of identity with the XlnR in *A. niger*. Homologs for this transcription factor are present in the *Trichocomaceae* family members, but not in yeast, basidiomycetes or zygomycetes. The AraR regulator has also been characterized in *A. nidulans* (Battaglia *et al.*, 2011b).

The disruption of the *araR* gene resulted in a reduced growth on arabinose, xylitol, arabinan, arabic gum, arabinogalactan and apple pectin in *A. niger* (Battaglia *et al.*, 2011a). But also, a reduced expression level of the arabinose and xylose metabolism genes was detected in the *A. niger* Δ *araR* when it is growing in arabinose (Battaglia *et al.*, 2011a; de Souza *et al.*, 2013). Both, AraR and XlnR regulate the pentose catabolism pathway in a coordinated manner. In *A. nidulans*, *araR* deletion nearly impaired the growth on arabinose and slightly reduced it on arabitol (Battaglia *et al.*, 2011b). However, in *A. nidulans*, the expression of some of the Pentose Catabolic Pathway (PCP) genes are induced by arabinose, but independently from AraR, which suggests that another unknown regulator might be involved (Battaglia *et al.*, 2011b).

1.3.3. The CLR1 and CLR2 regulators

Coradetti *et al.* (2012) described two new transcription factors, CLR1 and CLR2, that are required for the degradation of cellulose in *N. crassa*. The CLR1 transcription factor is induced under cellulolytic conditions and it is involved in the regulation of the genes coding for plant polymers degradation, in particular the β -glucosidase genes. The transcription factor CLR2, described in the same fungus, is expressed under the control of CLR1 and directly induces cellulase and some hemicellulase genes (Coradetti *et al.*, 2012). The *N. crassa* Δ *clr2* mutant showed no cellulase activity on Avicel, and had only trace levels of hemicellulase activity.

The *clr2* gene is the ortholog of the *A. oryzae manR* (Ogawa *et al.*, 2012b) and *A. nidulans clrB* (Coradetti *et al.*, 2013). ManR from *A. oryzae* is a transcription factor of 771 amino-acids. The gene *manR* encodes 2 different domains: the Zn(II)₂Cys₆ binuclear cluster domain (Pfam00172) and the fungal specific transcription factor domain (Pfam04082) (Ogawa *et al.*, 2012b). The deletion of *manR* caused a dramatic decrease in the activity of the β -mannanases in *A. oryzae*, and its overexpression caused an increased activity of 2.9 and 6.8-fold when exposed to konjac glucomannan and carob galactomannan, respectively. The same pattern was observed for the α -galactosidase activity. These results were confirmed by high throughput expression studies that determined that the expression of the genes coding for mannanolytic enzymes was significantly affected by the disruption of *manR* (Ogawa *et al.*, 2012b). The cellobiohydrolases encoding genes, as well as an endoglucanase gene and a β -glucosidase gene were shown to be controlled by ManR through array experiments (Ogawa *et al.*, 2012a). In *A. nidulans*, the expression of cellulases was importantly decreased in the Δ *clrB* mutant, whereas the genes coding for hemicellulases were induced when this strain is exposed to microcrystalline pure cellulose, also called Avicel. Genes involved in secondary metabolism showed a partial induction in the Δ *clrB* mutant exposed to Avicel, indicating that ClrB is necessary for their expression, but that other inducers are also involved (Coradetti *et al.*, 2013).

1.3.4. The Pal pathway

The Pal pathway regulates the gene expression in a pH-dependent fashion; at alkaline pH, it activates alkaline-expressed genes and represses acid-expressed genes. There are seven genes known to be involved in the signalling pathway that leads to the activation of the PacC protein, which induces or represses different genes (Peñalva *et al.*, 2008). The genes involved in the signalling pathway are *palH*, *palF*, *pall*, *vsp32*, *palB*, *palC* and *palA*. PalH is a 7-transmembrane domain protein, PalF is an arrestin like protein, Pall a 3-transmembrane domain protein that assists PalH in the plasma membrane localization, Vps32 (Snf7) binds to endosome membranes directly, PalB is a calpain-like cysteine protease and PalC and PalA both interact with Vps32. The interaction between the different proteins of this pathway is represented in Figure 1.12.

PacC is the transcription factor in this pathway. It contains three C₂H₂ zinc finger domains and needs to be proteolytically cleaved in order to be active. In acidic conditions, it is the full length form, PacC⁷², that predominates. When the pH rises PacC undergoes two forms of cleavages. The first one is carried out by the calpain-like cysteine protease PalB, which cleaves approximately 180 amino-acids from the C-terminal of the protein and leads to the PacC⁵³ form. The second one is a proteolysis catalysed by a proteasome that removes approximately 245 amino-acids from the C-terminal. This second cleavage leaves the binding domain intact and ready to interact with the DNA (Hervás-Aguilar *et al.*, 2007; Penas *et al.*, 2007).

There is evidence that this pathway regulates the expression of glycoside hydrolases encoding genes, such as the xylanases in *A. nidulans* (MacCabe *et al.*, 1998), or the arabinofuranosidases (Gielkens *et al.*, 1999).

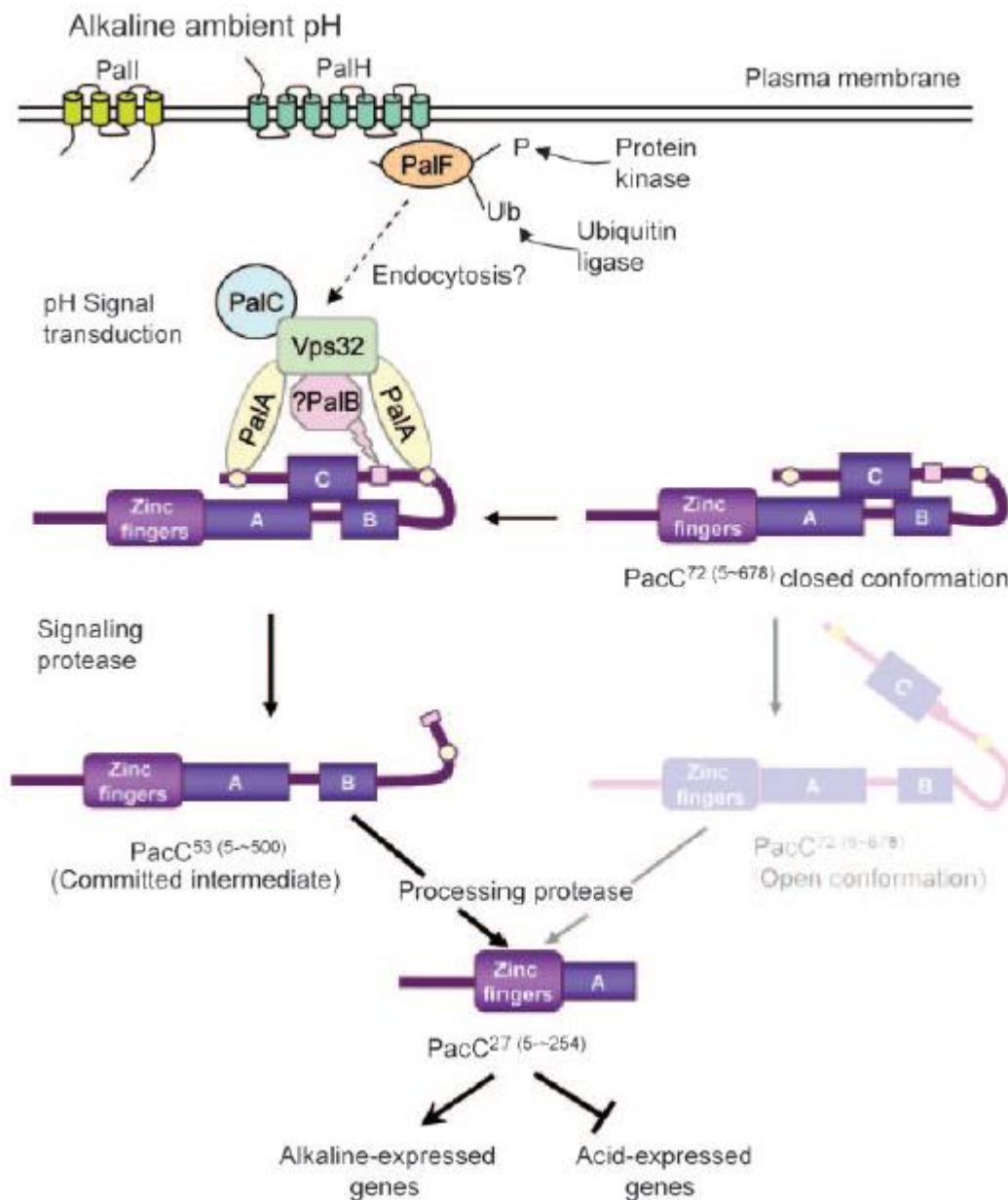


Figure 1.12: The pH regulatory pathway in filamentous fungi. Under acidic conditions, PacC is inaccessible to the protease. But, under alkaline conditions, the alkaline ambient pH signal is transmitted from the plasma membrane complex, by a process that probably involves endocytosis of PalH and/or PalIF and the participation of PalC, to the endosomal membrane complex. Here, PacC, binds by PalA and undergoes a proteolysis. The proteolysis is probably mediated by the protease PalB, which removes the negatively acting C terminus to form PacC⁵³. PacC⁵³ has an open conformation that is accessible to the processing protease, the proteasome, which removes a further 245 residues in a pH-independent manner to form the PacC²⁷ activator of alkaline-expressed genes and repressor of acid-expressed genes. Taken from Penas *et al.* (2007).

1.3.5. The ACEI and ACEII transcription factors

The ACEI transcription factor was described by Saloheimo *et al.* (2000) as an activator of the cellobiohydrolases genes in *T. reesei*, by the same kind of screening that revealed the existence of ACEII. The screening consisted in the selection of transcriptional activators from a cDNA library expressed in yeast, by their promoter binding and activation properties. This transcription factor contains a C₂H₂ zinc finger domains and binds to the core sequence 5' AGGCA 3'. The deletion mutant presented a reduced growth on cellulose plates, suggesting that this transcription factor regulates the expression of cellulase genes.

However, further studies on the ACEI regulator revealed that the gene cloned originally lacked the N-terminal region and that the full length ACEI is actually a repressor of the expression of cellulase and xylanase genes in *T. reesei*. It was proposed that the truncated version of ACEI has a different regulatory function than the full length protein. The deletion of *aceI* not only produced a reduced growth on sorbitol, but also an increase in cellulase and xylanase genes expression levels (Aro *et al.*, 2003).

The ACEII regulator was described in *T. reesei*. This is another regulator with a Zn(II)₂Cys₆ binuclear cluster domain that shares the core binding sequence 5' GGCTAA 3' with the XlnR/XYRI regulator (Aro *et al.*, 2001). It has been suggested that phosphorylation and probably heterodimerization of the two transcription factors, XYRI and ACEII in *T. reesei*, may be necessary for the binding to the promoter region (Stricker *et al.*, 2008). The *T. reesei* deletion mutant in the *aceII* encoding gene produced less cellobiohydrolases and endoglucanases than the wild type strain when it was exposed to cellulose. However, the induction of the expression of these genes by sophorose is not affected by the lack of the ACEII transcription factor (Aro *et al.*, 2001).

It is important to note that no homolog to the ACEII transcription factor was found in any Aspergilli, or other fungi such as *N. crassa*, *F. graminearum*, etc. This suggests that it is a characteristic regulatory system to *Trichoderma* and it is not widespread amongst fungi, but this also

implies that other transcription factors are used by these fungi to achieve the induction of the target genes (Aro *et al.*, 2005; Stricker *et al.*, 2008).

1.3.6. The AreA regulator

The AreA regulator belongs to the C₄ zinc finger domain class. It is involved in the use of alternative nitrogen sources, like nitrate, nitrite or purines. Hence, it controls the expression of the nitrogen catabolism genes under nitrogen limiting conditions (Aro *et al.*, 2005). An *A. nidulans* deletion mutant in the *areA* locus was shown to have a reduced cellulases secretion compared to the wild-type strain, suggesting that the cellulase-encoding genes are directly or indirectly regulated by AreA (Lockington *et al.*, 2002).

1.3.7. The carbon catabolite repression system

The carbon catabolite repression, also called CCR, is probably the most well-known repressing system controlling the expression of the GH-encoding genes and other genes involved in the utilisation of alternative carbon sources in filamentous fungi. This system has been studied in an important number of microorganisms, in particular in the yeast *S. cerevisiae*, where the Mig1 protein represses the transcription of genes associated to the utilisation of alternative carbon sources (Santangelo, 2006). The presence of an easily metabolisable carbon source that is also energetically convenient results in the repression of an important set of genes. In filamentous fungi the repression is mediated by the CreA regulator. CreA is a C₂H₂ zinc finger regulator and a mutation in the gene coding for this protein leads to a de-repression of the cellulase and xylanase genes (Aro *et al.*, 2005; de Vries & Visser, 2001). For instance, the hypercellulolytic *T. reesei* Rut-C30 strain carries a truncated *cre1* gene, and it can therefore induce the expression of GH-encoding genes while growing with glucose as carbon source (Ilmén *et al.*, 1996). Nevertheless, the full deletion of *creA* is a modification not always easy to achieve: the work done by Shroff *et al.* (1997) showed that the partial deletion of *creA* could be achieved in *A. nidulans*, but different modifications lead to different

phenotypes and not always a complete loss-of-function is attained. They demonstrated that in this fungus an intact 3' flanking region of the locus is necessary to get a viable strain.

Interestingly, Tamayo *et al.* (2008) showed that CreA regulates the expression of xylanase genes in *A. nidulans* through the control of the *xlnR* encoding gene expression. In fact, CreA directly represses the expression of *xlnR*. In absence of CreA the *xlnR* gene is de-repressed and the XlnR protein will strongly activate the xylanase encoding genes. These findings are in accordance with the work of Mach-Aigner *et al.* (2008) in which a similar regulatory mechanism of the expression of *xyr1* was described for *T. reesei*. On the other hand, Sun & Glass (2011) determined the regulon of CRE1 in *N. crassa* using full genome arrays and the direct targets of CRE1 were determined using CHIP-PCR. The direct targets of CRE1 in *N. crassa* include the *amyR* regulator, involved in the starch degradation, a cellobiohydrolase, two xylanases and a monosaccharide transporter. However, a set of 271 genes whose expression was induced in absence of CRE1 was identified, which included the direct targets of CRE1, transporters, carbohydrate metabolism genes, amongst others. Also, 381 genes whose expression was repressed in the $\Delta cre1$ strain were detected.

It is important to note that there are two additional proteins involved in the carbon catabolite repression system in fungi: CreB and CreC. The interaction between these two proteins and CreA is shown in Figure 1.13. CreB is a functional deubiquitinating enzyme (Lockington & Kelly, 2001) and CreC contains a five WD40 repeat motifs and a proline-rich region that might be associated with protein-protein interactions (Todd *et al.*, 2000). Lockington & Kelly (2002) showed that both CreB and CreC interact and are involved in the CCR response in *A. nidulans*, but also that CreB acts downstream of CreC. CreB contains four PEST regions which are strongly implicated in the signalling for proteolysis and rapid degradation. It has been suggested that the interaction between CreB and CreC masks the PEST regions present in CreB and stabilizes the protein in both, repressing and de-repressing conditions. Furthermore, in repressing conditions, free CreB removes the ubiquitin moieties from CreA, which can now repress the expression of its target genes. In de-repressing

conditions, CreB is degraded, therefore the ubiquitin moieties of CreA are not removed and the protein is targeted for degradation via proteasome; without CreA the de-repression of the target genes occurs. In accordance with the previous model, the deletion of the gene coding for the deubiquitinating protein CreB in *A. oryzae* (Hunter *et al.*, 2013) and *T. reesei* (Denton & Kelly, 2011) increased the expression of hydrolase encoding genes in these fungi. Similarly, the deletion of both *creA* and *creB* increased the production of α -amylase in *A. oryzae* (Ichinose *et al.*, 2014).

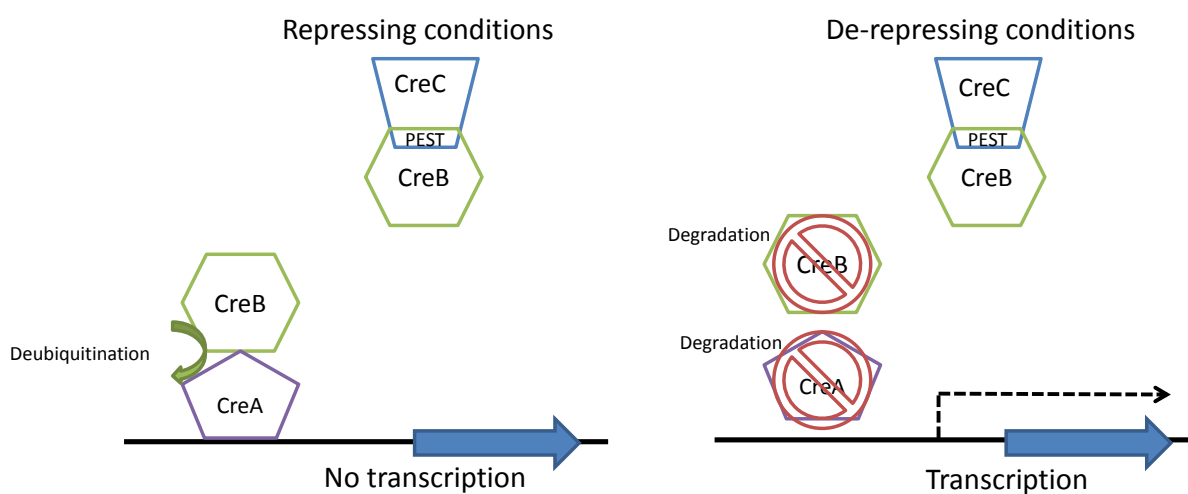


Figure 1.13: Carbon Catabolite Repression model. In repressing conditions (left panel), CreB is present and it removes the ubiquitin moieties from CreA, resulting in high levels of CreA, hence the genes controlled by CreA are repressed. In carbon de-repressing conditions, free CreB is degraded via its PEST sequences; therefore the ubiquitin moieties from CreA are not removed. Then, CreA is targeted for degradation via the proteasome; the promoters of the genes under its control are released and de-repressed. In both repressing and de-repressing conditions, some CreB is present in a complex with CreC, and this is required for either the expression or the stability of the enzymes and permeases required for the use of, for example, proline. Modified from Lockington & Kelly (2002).

1.3.8. The Unfolded Protein Response and the HacA transcription factor

The presence of misfolded proteins in the secretion pathway stimulates the UPR, which serves to alleviate the overloaded secretion system and help the correct folding of the proteins that are being processed. In *T. reesei* and *A. nidulans* the unfolded proteins are sensed by a transmembrane

riboendonuclease, IRE1, which oligomerises and is subsequently activated by auto-phosphorylation. The *hac1/A* mRNA is the substrate for IRE1, from which a 20 nucleotides intron is removed and the rest of the RNA is joined together by a tRNA ligase. Additionally, 200 nucleotides at the 5' end of the mRNA are truncated. In these fungi, a double mechanism is necessary for the induction of *hac1/A*: the splicing that restores the correct reading frame, and the use of an alternative transcription start site that omits the upstream ORF that inhibits the use of the correct translation start site (Geysens *et al.*, 2009; Saloheimo *et al.*, 2003). The activated *hac1/A* mRNA is then translated to the Hac1/A transcription factor, which mediates the UPR.

The Hac1/A transcription factor is supposed to bind to UPR elements within the promoter regions of its target genes. The UPR affects transcription of a few hundred genes involved in processes such as protein translocation, protein glycosylation, lipid and inositol metabolism and ER homeostasis (Saloheimo *et al.*, 2003). In *A. niger*, HacA is able to up-regulate its own transcription via HacA binding sites in its own promoter region (Mulder & Nikolaev, 2009; Mulder *et al.*, 2004). In *T. reesei* down-regulation of genes encoding proteins destined for extracellular secretion is also observed but it is unknown whether this response is directly linked to the UPR (Arvas *et al.*, 2006). It has been proposed that in *A. niger* the ER stress can lead to the down-regulation of genes encoding secreted proteins; this effect is known as the repression under ER stress (RESS) and it may be independent from the UPR (Guillemette *et al.*, 2007). It was suggested that the response to secretion stress, may actually be a more general mechanism not only to alleviate accumulation of unfolded proteins in more extreme conditions, but also to modulate secretion capacity in response to the conditional need for extracellular enzymes when the fungus is not exposed to a chemically induced stress (Jørgensen *et al.*, 2009).

1.3.9. Noncoding RNAs and Antisense transcription

There are different kinds of noncoding RNAs: the housekeeping noncoding RNAs include ribosomal, transfer and small nuclear RNAs and are usually expressed constitutively. There are also

short regulatory noncoding RNAs: microRNAs, small interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs), all constitute the RNA interference (RNAi) pathways. And finally, there are the long noncoding RNAs, also called lncRNAs (Ponting *et al.*, 2009).

siRNAs and miRNAs are both derived from double-stranded RNA (dsRNA) precursors that are recognized and processed by Dicer to generate short duplexes. On the one hand, the siRNAs are derived from exogenous dsRNA, like viral RNA, or endogenous transcripts from repetitive sequences (*e.g.*, transposable elements), or from transcripts that can form long hairpins. This kind of RNAs usually fully matches their mRNA targets and triggers the RNA cleavage or transcriptional silencing mediated by the Argonaute proteins. The miRNAs, on the other hand, are generated from miRNA-encoding genes that originate ssRNA precursor transcripts that form hairpin structures. The miRNAs can target mRNAs that are not fully complementary and cause mRNA degradation and translational repression (Dang *et al.*, 2011).

The long noncoding RNAs may be located within the nucleus or cytosol, may or may not be polyadenylated, and are often transcribed from either strand within a protein-coding locus. There are 5 categories of lncRNAs: (1) sense, or (2) antisense, when overlapping one or more exons of another transcript on the same, or opposite, strand, respectively; (3) bidirectional, when the expression of it and a neighbouring coding transcript on the opposite strand is initiated in close genomic proximity; (4) intronic, when it is derived solely from within an intron of a second transcript; (5) intergenic, when it lies within the genomic interval between two genes (Ponting *et al.*, 2009). The natural antisense transcripts (NATs) are a subset of the lncRNAs and are RNA molecules transcribed from the opposite DNA strand to other transcripts and overlap in part with sense RNA. They can be divided in two groups: *cis*- and *trans*-NATs. The first ones are transcribed from the same genomic region as their complementary sense RNA, and the *trans*-NATs are complementary to transcripts that come from a remote genomic locus. Their mechanisms of action can be broadly divided into four categories: 1) transcription-related, 2) RNA-DNA interactions, 3) RNA-RNA interactions in the nucleus and 4) RNA-

RNA interactions in the cytoplasm (Figure 1.14). Cytoplasmic sense–antisense RNA duplexes can hide or expose AU-rich elements in sense transcripts, affecting RNA stability. They can provoke changes in the RNA secondary structure, alter translation, subcellular localization and accessibility to the RNA degradation machinery. Also, they can activate the RNA interference (RNAi) pathway (Donaldson & Saville, 2012; Faghihi & Wahlestedt, 2009). Lately, the high-throughput sequencing technologies have emphasised the existence of lncRNAs and NATs in a range of eukaryotic organisms, including the filamentous fungi *N. crassa*, *T. reesei*, *A. nidulans* and *A. niger* (Arthanari *et al.*, 2014; Delmas *et al.*, 2012; Ries *et al.*, 2013; Sibthorp *et al.*, 2013). It has been shown, for instance, that in *N. crassa* the antisense transcript of the *frq* gene influences the circadian timing, since the removal of the antisense promoter modifies the phase of the rhythm (Kramer *et al.*, 2003).

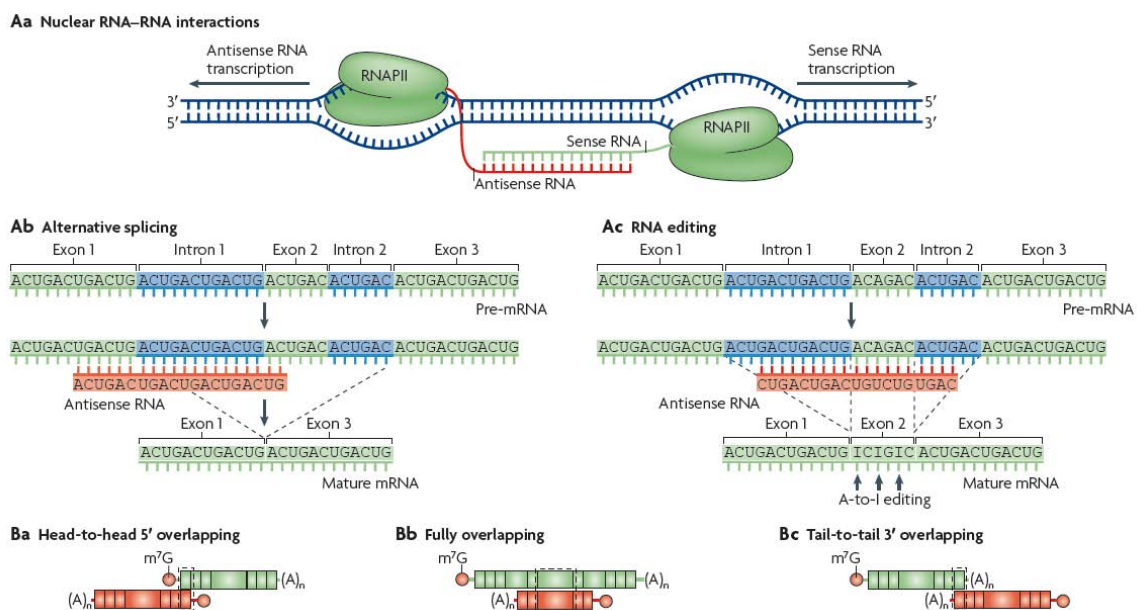


Figure 1.14: Nuclear and cytoplasmic sense–antisense RNA pairing. A) Nuclear RNA duplex formation might occur locally after transcription, thereby inhibiting sense RNA processing (Aa). Natural antisense transcripts can cover donor and acceptor splice sites in the sense pre-mRNA transcript to change alternative splicing patterns (Ab). RNA editing is another possible consequence of nuclear RNA duplex formation, for example A-to-I editing by the ADAR (adenosine deaminase that act on RNA) enzyme (Ac). RNA editing can induce nuclear retention and degradation of hyper-edited transcript by inosine-specific nucleases or can result in alteration of the amino acid sequence. B)

Cytoplasmic sense–antisense duplex formation can occur in three different formats: the head-to-head 5' overlapping configuration (Ba), the fully overlapping configuration, in which the antisense transcript is embedded in the sense transcript (Bb), or the tail-to-tail 3' overlapping configuration (Bc). Exons are depicted as boxes, and the overlapping region is shown with a dashed border. Taken from Faghihi & Wahlestedt (2009).

RNA interference (RNAi) is a conserved eukaryotic gene silencing mechanism that has several functions: genomic defence, heterochromatin formation and gene regulation. When the RNAi pathway is activated it results in the degradation of transposon RNA, viral RNA or any other mRNA. The RNAi machinery consists of three core components: Dicer, Argonaute and RNA-dependent RNA polymerase (RdRP). The RdRP can generate a dsRNA from ssRNA. In response to the presence of the dsRNA, the RNA polymerase III enzyme, called Dicer, processes the dsRNA into small interfering RNAs of ≈25 nucleotides (siRNAs). One of the RNA strands is then incorporated into the RNA-induced silencing complex (RISC) whose core catalytic component is an Argonaute protein. One of the strands of the dsRNA, loaded onto RISC, needs to be removed in order to activate the Argonaute protein subunit of this complex and uses the remaining single-stranded small RNA as a guide to silence the target RNAs (Li *et al.*, 2010; Tijsterman & Plasterk, 2004).

1.4. Tools for the study of the genome and the transcriptome

1.4.1. Genome sequencing technologies

Depicting the DNA sequences was made possible by two main techniques: the degradation method introduced by Maxam & Gilbert (1977) and the chain termination sequencing method designed by Sanger *et al.* (1977). Both techniques were developed in the late 1970s, and even though they were difficult and laborious, they represented the beginnings of the genomics as we know it today. In fact, the Sanger's method proved to be more useful and was the basis for the development of other more effective sequencing technologies. Lately, the development of sequencing technologies has been extremely rapid, and the techniques are upgraded constantly.

Chapter I

The next generation sequencing (NGS) technologies constitute various strategies that rely on a combination of template preparation, sequencing and imaging and genome alignment and assembly methods. These techniques have the capacity to deliver a great amount of information in a very short time and at an affordable price (Metzker, 2010). Moreover, the application of such NGS is not limited to the sequencing of DNA, but transcriptomes or chromatin immunoprecipitation (ChIP) are also the object of sequencing (Fullwood *et al.*, 2009).

One of the most important technique is probably the pyrosequencing, which is a method designed to detect the incorporation of a nucleotide into a growing DNA strand by bioluminescence. This method was able to complement the Sanger's method and was developed into a platform now known as Roche/454 (Stranneheim & Lundeberg, 2012). In pyrosequencing, as the nucleotides are incorporated by a polymerase, pyrophosphate and protons are generated. The pyrophosphates are used in a cascade of enzymatic reactions that ultimately generates a light burst thanks to the luciferase that generates the photons that are detected and recorded as flowgrams (Metzker, 2010; Stranneheim & Lundeberg, 2012)(Figure 1.15). Protons generated during the incorporation of the nucleotide can also be detected as a change in pH; this is the principle of the latest pyrosequencing technology, Ion Torrent (Life Technologies) (Steen & Cooper, 2011).

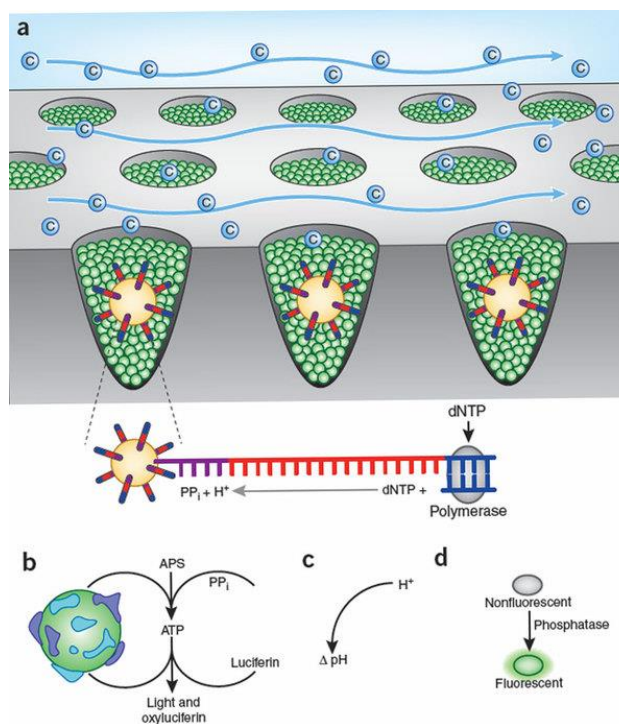


Figure 1.15: Representation of the pyrosequencing technology. a) Beads covered with amplified DNA are deposited into a microreactor, dNTPs are sequentially added and washed over the wells. The incorporation of a nucleotide triggers an enzymatic reaction that produces light (b), protons can be detected as a change in pH (c) or a phosphatase cleaves the phosphate from the fluorophore (d). APS, adenosine 5' phosphosulfate. Taken from Steen & Cooper (2011).

Another technology that became one of the most used nowadays is the Illumina/Solexa. In this case, the template DNA is ligated to an adaptor which is fixed to a surface. Each DNA template is amplified by solid-phase PCR, creating a surface covered with a high density of spatially distinct clusters, each of which contains a single DNA template. The DNA templates are sequenced by passing the four spectrally distinct reversible dye terminators in a flow solution over the surface in presence of the DNA polymerase. Only one nucleotide can be incorporated, thanks to the 3' modification of the nucleotides. The incorporation of a nucleotide in a cluster is detected by fluorescence imaging. After the imaging the dye terminator is removed and the cycle can start once again (Stranneheim & Lundeberg, 2012) (Figure 1.16).

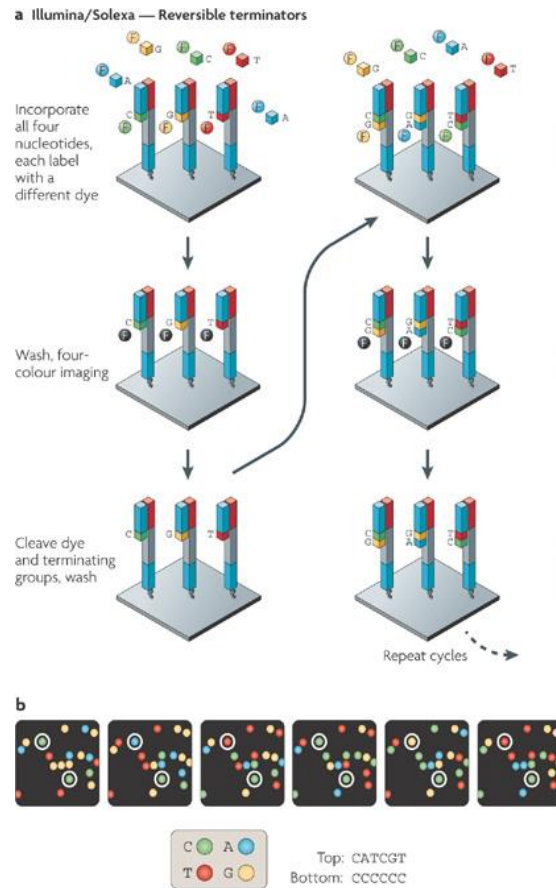


Figure 1.16: Representation of the Illumina/Solexa technology. In the first step, the labelled nucleotides are incorporated, the non incorporated ones are washed and the imaging can take place. The cleavage of the fluorescent dye follows and a new cycle can begin. Taken from Metzker (2010).

The sequencing by ligation technology, such as the one used by the SOLiD platform, is different from the pyrosequencing in that it uses a ligase. During the sequencing reaction, octamers (oligomers of eight nucleotides) are used as probes, and they contain two known bases at the 3' end followed by three degenerate nucleotides and three universal nucleotides that can hybridize any other nucleotide. Four different fluorescent labels are attached to the probes, each label is attached to four different octamers; thus there are 16 different octamers that cover all the possible combinations of the two known bases. The probe fluorescently labelled hybridizes to its complementary sequence adjacent to the primer template. The DNA ligase joins the labelled probe to the primer and the non-ligated probes are washed away. Fluorescence imaging allows determining the identity of the

hybridized probe. This cycle is repeated ten times. And a second ligation round starts, but with a primer n-1, which switches the interrogation bases and the corresponding ten colours one position to the left. Ten cycles follow, and 3 more rounds of ligation. The colour calls from the 5 rounds of sequencing are ordered into a linear sequence to depict the nucleotide sequence underlying (Metzker, 2010; Stranneheim & Lundeberg, 2012) (Figure 1.17). The sequencing by ligation technology produces reads of 50 bp, which is very short compared to the ≈ 500 bp that the Roche/454 can deliver with the pyrosequencing method. However, the fact that the bases are read several times, provides an inherent error correction system, and the SOLiD technology becomes more suitable for projects in which the error rate needs to be low.

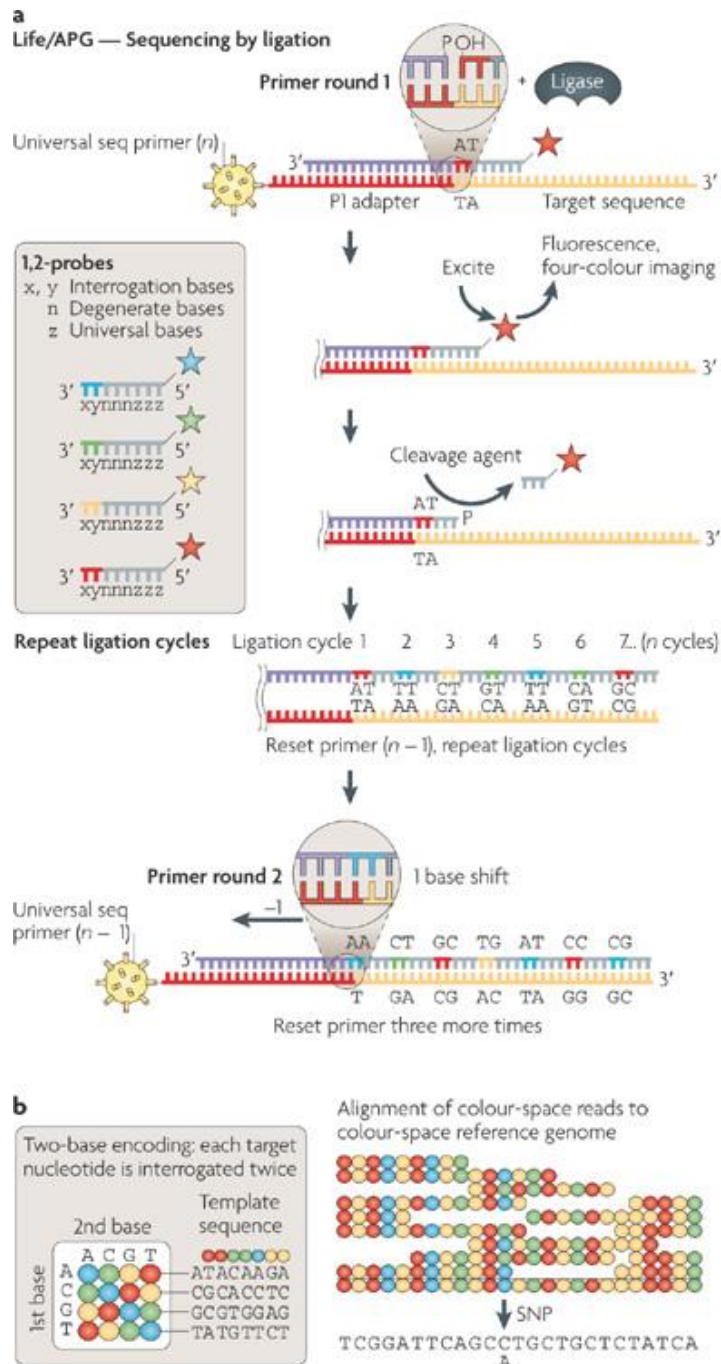


Figure 1.17: Representation of the SOLiD technology. a) 16 different octamers are used in this reaction. The correct probe will hybridize the complementary sequence and a ligation follows. The ligated probes are cleaved and the colour of the probe can be imaged. The cycle is repeated 10 times. Four more ligation cycles are performed each of which will shift the sequencing one position to the left. b) Four different dinucleotides are associated with one colour (for example, AA, CC, GG and TT are coded with a blue dye).

The bases are interrogated twice and the compiled data generates a colour-space data bits which can be decoded into a DNA sequence. Taken from Metzker (2010).

These methods, amongst others, have revolutionized the genomics. However, they all depend on the PCR amplification of the DNA template, which can introduce a bias, since there are regions of the DNA that can be more easily amplified than others. The solution to this is the single-molecule sequencing that eliminated the PCR amplification step and minimises the amount of input DNA. This kind of method was applied by Helicos Biosciences and Pacific Biosciences.

Once the sequences have been obtained, it is necessary to assembly them into contigs, scaffolds and, when it is possible, chromosomes, depending on the quality of the sequencing and the number of gaps left. Finally, the annotation step aims to predict the possible ORFs encoded, with their intron-exon structure. Even though the public databases are used to predict the genes, mistakes can be made in this step, and this is the reason why a manual curation of the annotation is always necessary.

1.4.2. Reference genes for RT-qPCR data normalisation

Better knowledge of gene function usually begins by investigating expression of the genes of interest (GOIs) under a broad set of culture conditions. Several techniques have been developed over the years to measure expression levels, including Northern blot, cDNA microarrays and RNA-seq. Yet, the coupling between reverse transcription and quantitative PCR (RT-qPCR) appears to be the most appropriate to study limited numbers of genes in large sets of conditions (Bustin *et al.*, 2005; Wong & Medrano, 2005). This technology uses the quantitative relationship between the amount of target present in a sample at the beginning of the reaction and the amount of PCR product at the moment when amplification of a PCR product exceeds the background noise and is first detected. The fewer cycles it takes to reach a detectable level of fluorescence, the greater the initial copy number of the target (Huggett & Bustin, 2011). Significant technical advances made this mRNA quantification method very accessible, highly specific and sensitive, but numerous critical issues in the workflow

need to be addressed before biologically meaningful and trustworthy conclusions can be drawn (Derveaux *et al.*, 2010).

The establishment of the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines is a helpful tool to design the experiment, to keep track of the experimental data in each of the quantification steps and to improve analysis post-quantification (Bustin *et al.*, 2009; Huggett & Bustin, 2011). Importantly, and no matter what the technique for measuring gene expression is, data normalisation is a critical step.

The normalisation process helps to control the variation in the total mass of RNA analysed and corrects the bias introduced by the biological sampling and techniques used in the whole process of quantification. Without normalisation, both of these pose serious problems for gene expression analyses and can either mask or overstate true expression levels of the genes of interest. An appropriate normalisation of each step would be ideal. The first step to be normalised is the sample size. It is important that the samples compared in an experiment have the same size; this ensures that the differences in efficiencies in the RNA extraction are avoided. The second normalisation point is ensuring that the same amount of RNA is used for every reverse transcription reaction. The RNA quality is essential in this phase. It would be desirable to measure the cDNA reverse transcribed from RNA as another normalisation point. The last, but not the least, is the use of reference genes. These genes represent the most common normalisation method for RT-qPCR data. It consists in targeting RNAs that are, it is hoped, universally and constitutively expressed, and whose expression do not differ between the experimental conditions. They are used to report any variation that occurs due to experimental error (Dorak, 2007). Another normalisation techniques that has been proposed is the use of external controls (Devonshire *et al.*, 2010; Liu & Slininger, 2007), a normalization strategy stimulated by the ERCC or EQUAL-quant programs (Baker *et al.*, 2005; Ramsden *et al.*, 2006), and especially relevant for the assessment of technical robustness in clinical and biological diagnostic laboratories. But normalization of gene expression levels by reference genes (internal controls) is

most certainly the gold standard (Hruz *et al.*, 2011; Kozera & Rapacz, 2013; Thellin *et al.*, 2009; Vandesompele *et al.*, 2009).

It is now clearly established that the use of a single reference gene is no longer acceptable as there is not a single gene that has a stable expression over all kinds of culture conditions or among different cell types. More than ten years ago, Vandesompele *et al.* (2002) first developed a method called geNorm, which allows the evaluation and ranking of candidate reference genes in terms of expression stability, or suitability as normalising gene. In a subsequent step, the algorithm is able to indicate how many reference genes are optimally required to remove most of the technical variation. Hence, good practice in data normalization for gene expression analysis relies on the identification, experimental validation and use of several reference genes.

In filamentous fungi based projects, such efforts have been observed with recent publications dedicated to the validation of suitable reference genes under specific experimental contexts (Bohle *et al.*, 2007; Fang & Bidochka, 2006; Hacquard *et al.*, 2011; Huang Xueling *et al.*, 2012; Raffaello & Asiegbu, 2013; Steiger *et al.*, 2010; Vieira *et al.*, 2011; Yan & Liou, 2006; Zampieri *et al.*, 2014; Zhang *et al.*, 2009; Zhou *et al.*, 2012). In Zhou's work (Zhou *et al.*, 2012), genes like *cypB* and *crzA* were evaluated because of their stability in transcriptomic datasets. Similarly, Kim & Yun (2011) selected 8 reference genes from transcriptomic data available on *Fusarium graminearum*. Such an ideal approach was an exception, as most often, authors have evaluated more classic "housekeeping genes" encoding for example the actin, the glyceraldehyde-3-phosphate dehydrogenase or the β -tubulin, which are still and too frequently used as single, non-validated reference genes.

1.4.3. Molecular tools for the genetic modification of filamentous fungi

The genetic engineering techniques of filamentous fungi are powerful tools for the study of the biology of these organisms, but also for the improvement of production levels of the desired metabolite, or the production of heterologous compounds. But the development of efficient

methods for the genes deletion or modification in filamentous fungi is still limited. Several transformation techniques have been proposed, although their efficiency depends on each particular species, therefore predictions on the feasibility of a method are difficult to make (Meyer, 2008).

The transformation itself can be achieved in different ways: the *Agrobacterium* mediated transformation (AMT) or the protoplasts mediated transformation are commonly used techniques in filamentous fungi. Other techniques have also been developed, such as electroporation and biolistic transformation (Jiang *et al.*, 2013; Rivera *et al.*, 2014).

The *Agrobacterium*-mediated transformation consists in the use of *Agrobacterium tumefaciens*, a plant pathogen that transfers part of its DNA located in the Ti plasmid, also called T-DNA, to the host cell. The plasmid contains a fragment with the *vir* genes that confer the virulence and include genes involved in the formation, transport and integration of the T-DNA. The sequences in the T-DNA can be replaced by the chosen DNA sequences without any negative effect (Michielse *et al.*, 2005). This system was first used for plant transformation, but it was then adapted to other eukaryotes, like filamentous fungi. However, the efficiency of this technique is highly dependent on the fungi (de Groot *et al.*, 1998). The transformation is usually performed with the binary system, in which the T-DNA and the *vir* region are located in two separate plasmids (Frandsen, 2011). The advantage of this method is the possibility to use any kind of starting material: protoplasts, spores, mycelium or fruiting bodies, although the efficiency with the different materials depends on the fungal strain used. Some strains are not susceptible of being transformed by this method, like *Sclerotinia sclerotiorum* (Rolland *et al.*, 2003), and others, like *A. niger*, give non reproducible results (Michielse *et al.*, 2005).

The protoplast mediated transformation, on the other hand, is a laborious and time-consuming technique and can be troublesome for some fungi. Conidia or mycelium can be treated with cell wall degrading enzymes in order to obtain the protoplasts for the transformation. The enzymes used are not always similarly efficient, and tests to each batch of enzymes could be done. The protoplasts

need to be osmotically stabilized and the uptake of DNA needs the presence of polyethylene glycol. However, the preparation of protoplasts from fungal cells using degrading enzymes remains the most common method for preparation of cells for transformation (Rivera *et al.*, 2014; Ruiz-Díez, 2002).

Once the DNA has been imported into the cell, two different types of DNA integration can occur in fungi: the non-homologous end joining (NHEJ), which is the most common event; and the homologous recombination (HR), which occurs at very low frequency, about 1%. And two types of transformants can be distinguished: the knock-in, that carry a foreign DNA sequence in the target gene; and the knock-out, where the target gene is completely substituted by a marker (Kück & Hoff, 2010). The main problem is that the HR is the most useful for targeted genetic modification, but its frequency is very low and it is easier to have a NHEJ event instead. To circumvent this, NHEJ-inactivated strains are often used. The deletion of the *ku80* or *ku70* homolog genes originates strains that have a much higher frequency of homologous recombination. This kind of strains were constructed in several filamentous fungi and their use allowed to increment the gene targeting efficiency to more than 90% (Honda *et al.*, 2011; Krappmann *et al.*, 2006; Meyer *et al.*, 2007; Nayak *et al.*, 2006; Ninomiya *et al.*, 2004; da Silva Ferreira *et al.*, 2006; Takahashi *et al.*, 2006). However, it was demonstrated that the strains with an NHEJ deficient system present genomic instability (Zhang *et al.*, 2011a) and increased sensitivity to different chemicals (Ninomiya *et al.*, 2004). Nielsen *et al.* (2008) proposed a method for *A. nidulans* in which it is possible to re-establish the NHEJ system after the genetic modification was achieved, the same system was then adapted to the citric acid producing strain *A. niger* (Honda *et al.*, 2011). In this system, a transitory disruption of the *ku80* homolog is achieved by integrating a counter-selectable marker in the *ku80* locus. After all the genetic modifications are achieved, the *ku80* gene can be restored by direct repeat recombination.

When using the knock-out technique, the transformations are usually done with a linear DNA fragment that contains the selection marker flanked by the upstream and downstream regions of the target gene. But this method is limited to the availability of selection markers and strains with

different auxotrophies. The use of this method is then very limited. And it would be difficult to construct strains with more than one knock-out gene. A method to perform knock-outs with a marker that can be recycled was first described for *S. cerevisiae* (Erdeniz *et al.*, 1997) and then adapted to *A. nidulans* (Nielsen *et al.*, 2006). It consists on the construction of two different cassettes that contain the targeting sequence and a part of the selectable marker gene, but none of the two cassettes contain a full length selectable marker gene, as it is shown in Figure 1.18. When the homologous recombination between the target gene and the two DNA fragments occurs, a full length and functional marker is generated, and the transformants can be isolated in selective media. The DNA fragments were designed in such a way that a direct repeat sequence on each side of the marker gene is formed once the integration was done. Hence, the marked gene can be excised by direct repeat recombination, and this allows several rounds of gene targeting using the same selectable marker.

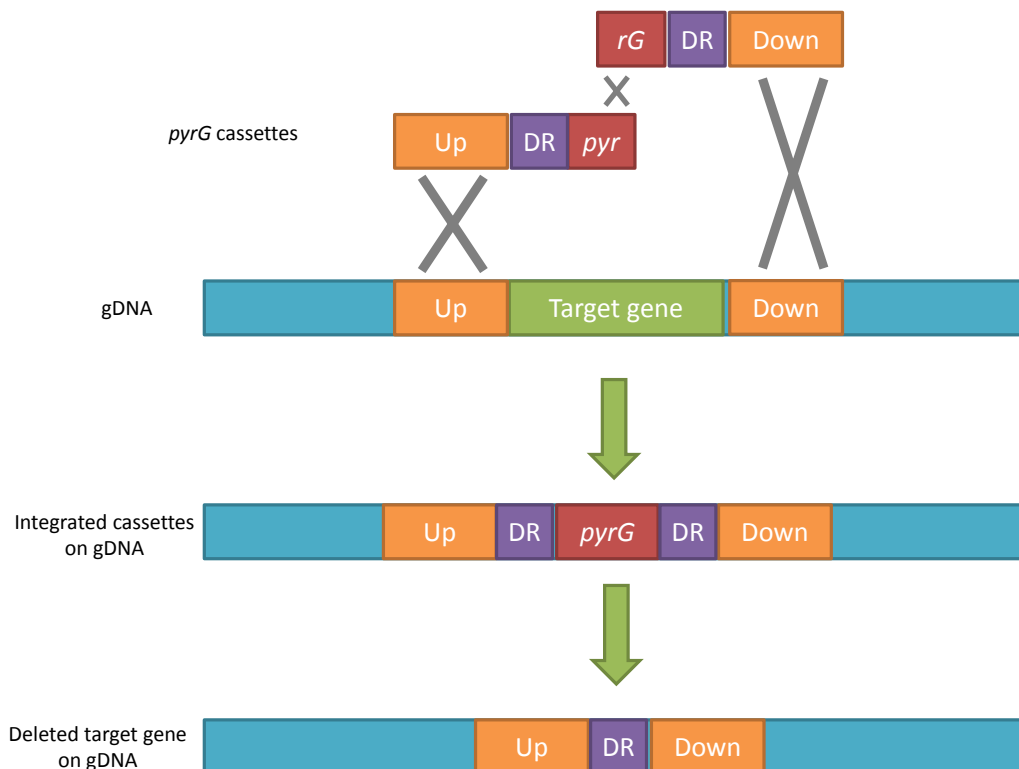


Figure 1.18: Schematic representation of the gene knock-out system with a bipartite selectable marker. Recombination between each fragment and the homologous

chromosomal locus, via the two targeting sequences up and down (orange boxes), results in deletion of the target gene and insertion of *pyrG* (red box) flanked by a direct repeat (violet boxes). After subsequent direct repeat recombination, the *pyrG* marker is excised leaving only a single DR sequence and the strain is ready for a new round of *pyrG* based gene-targeting. Modified from Nielsen *et al.* (2006).

2. General materials and methods

2.1. Microbiology techniques

2.1.1. Chemicals and Reagents

Most of chemicals used in this project were obtained from Sigma and were of analytical grade or molecular biology grade. Other suppliers are referred to in the text. All solutions were made up in ultrapure H₂O and autoclaved at 121°C for 20 minutes. Millipore 0.22 µm filters were used to filter-sterilise the solutions that cannot be autoclaved.

2.1.2. Strains

The industrial strain used in this work, *Talaromyces versatilis* (IMI378536), is an ADISSEO proprietary strain (patent no. W0 99/57325). This strain derives by mutagenesis from the *T. versatilis* wild type strain, IMI134756. This strain was used to construct the *T. versatilis* Δ pyrG, the *T. versatilis* Δ clrA and the *T. versatilis* Δ xlnR mutants (see below). *Aspergillus niger* 8.1, *Aspergillus nidulans* 2.137 and *Trichoderma reesei* QM6a strains were kindly provided by David Archer's team.

2.1.3. Culture conditions

All agar slopes were made by pouring 10 mL of molten agar into 30 mL Universal bottles (Sterilin) and leaving the agar to set at an angle. Spores of *T. versatilis* were obtained by growing the strain on Potato Dextrose Agar (PDA) slopes for 7 to 10 days at 30°C. Once conidia were present they were harvested by washing the surface of the slopes with 3 mL 0.01% (w/v) Tween 80. After vortexing the spore suspension, a 10 µL sample was removed and an average count of the number of spores was determined using a Thoma cell counting chamber (Marienfeld). The amount of conidia (*N*) present in the 16 medium squares that form the big central square was counted. This was then multiplied by 10⁴ and by the dilution factor (*f*) to obtain the concentration of conidia per mL of suspension:

Chapter II

$conidia/mL = N \times 10^4 \times f$. The volume of suspension needed to inoculate a culture with a specific number of spores (e.g. 2×10^5 spores/mL in 100 mL of TvMM) was calculated with the formula: $V_1 \times C_1 = V_2 \times C_2$, where V is the volume and C is the concentration. The spore suspensions were used to inoculate liquid medium, plates or slopes. Glycerol stocks were made by mixing 1 volume of the original spore suspension with 1 volume of 50% (w/v) glycerol in sterile cryotubes, which were stored at -80°C .

The *T. versatilis* minimal media (TvMM) contained for 1 L: 1.9 g KH_2PO_4 , 0.65 g KCl, 0.65 g MgSO_4 , 12.5 mg ZnSO_4 , 12.5 mg MnCl_2 , 12.5 mg FeSO_4 , 5 g NH_4Cl . The MM was supplemented with 10 g/L glucose as the sole carbon source, unless otherwise stated. The pH was adjusted to 6.0 with 50 mM KH_2PO_4 . The liquid medium was inoculated with 2×10^5 spores/mL in Erlenmeyer flasks. The cultures were carried out at 30°C and agitated at 150 RPM for 48 h. To prepare TvMM plates, the media was supplemented with 15 g/L of agar.

Aspergillus niger and *Aspergillus nidulans* were also kept on PDA. The spores produced in this media were counted as described before and used to inoculate *Aspergillus* minimal media (AMM) plates. The AMM contained for 1 L: 6 g NaNO_3 , 0.52 g KCl, 0.52 g $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$, 1.52 g KH_2PO_4 , 0.008 mg $\text{Na}_2\text{B}_4\text{O}_7 \cdot 10\text{H}_2\text{O}$, 0.16 mg $\text{CuSO}_4 \cdot 5\text{H}_2\text{O}$, 0.16 mg $\text{FePO}_4 \cdot \text{H}_2\text{O}$, 0.16 mg $\text{MnSO}_4 \cdot 4\text{H}_2\text{O}$, 0.16 mg $\text{NaMoO}_4 \cdot 2\text{H}_2\text{O}$, 1.6 mg ZnSO_4 . The AMM was supplemented with the appropriate carbon source to a final concentration of 1% (w/v). To prepare AMM plates, the media was supplemented with 15 g/L of agar.

Trichoderma reesei was kept on PDA as well. The spores of *T. reesei* were counted and used to inoculate the *Trichoderma* minimal media (TMM) plates, which contained per L: 10 g KH_2PO_4 , 0.005 g $\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$, 0.0016 g $\text{MnSO}_4 \cdot \text{H}_2\text{O}$, 0.0014 g $\text{ZnSO}_4 \cdot \text{H}_2\text{O}$, 0.0037 g $\text{CoCl}_{12} \cdot 6\text{H}_2\text{O}$, 0.6 g MgSO_4 , 0.6 g CaCl_2 , 15 g KH_2PO_4 and 5 g $(\text{NH}_4)_2\text{SO}_4$. The media was supplemented with carbon sources at a final concentration of 1% (w/v). To prepare TMM plates, the media was supplemented with 15 g/L of agar.

2.1.3.1. Growth on agar plates

T. versatilis MM (TvMM) supplemented with 15 g/L of agar was complemented with 1% glucose, 1% arabinose, 1% xylose, 1% Avicel, 1% xylan or 1% wheat straw. Media was poured into 90 mm plates and left to set. Plates were inoculated with a conidia suspension; each plate was inoculated with 100 conidia in the centre of the plate. Plates were incubated at 28°C. The mycelium growth was followed and photographed daily for 2 weeks.

A. niger and *A. nidulans* spores were used to inoculate AMM plates supplemented with 1% glucose, 1% arabinose, 1% xylose, 1% Avicel, 1% xylan or 1% wheat straw. A suspension containing 100 spores was deposited in the centre of each plate. Likewise, TMM plates supplemented with the same carbon sources were inoculated with *T. reesei* spores. Plates were incubated at 28°C. The mycelium growth was followed and photographed daily for 2 weeks.

2.1.3.2. Growth curves in liquid media

To generate *T. versatilis* growth curves, 50 mL cultures of AMM or TvMM complemented with glucose 1% as the carbon source were inoculated with 2×10^5 conidia/mL and incubated at 28°C and 150 RPM. Mycelia were harvested at the following time points: 43 h, 47 h, 51 h, 67 h, 71 h, 75 h and 92 h. Mycelia samples were separated from the culture liquid by pouring the content of each flask through Miracloth (Calbiochem® Merck). They were rinsed twice with sterile ultrapure H₂O before being blotted dry between paper towels. The mycelium was recovered and put into a pre-weighed Falcon tube. They were freeze-dried (SB4 model, HemLab England) overnight and biomass was weighed out the following day. Two biological replicates were done for each curve.

The average weight of biomass from the replicates for each time point was plotted (time versus weight). The doubling time (*dt*) was inferred from the graph by deducing the time it takes for the fungal biomass to double in mass (*e.g.* from 1 g to 2 g).

2.1.4. Harvest and storage of mycelia

Mycelia samples were collected by filtration using Miracloth (Calbiochem® Merck). They were rinsed twice with MM without any carbon source. For RNA extraction, the samples were put on 2 mL phase-lock tubes and flash frozen under liquid nitrogen. For DNA extraction, mycelia samples were placed in 1.5 mL tubes containing glass beads (Sigma) and flash frozen under liquid nitrogen. The mycelia samples were stored at -80°C. Mycelia samples filtered were also used to transfer to fresh media after being weighted.

2.1.5. Transformation of *Talaromyces versatilis*

Plasmids were transformed in *T. versatilis* $\Delta pyrG$ using PEG-mediated transformations with 1×10^7 protoplasts and 10 μg of plasmid DNA per reaction. For protoplasting, 6×10^8 conidia, inoculated into 200 mL of MN-Uri medium, were incubated for 17 h (30 °C, 150 RPM). Germinated conidia were harvested by centrifugation (2500 g, 4°C, for 20 min), washed with 50 mL cold 0.6 M MgSO_4 twice, then resuspended in 10 mL OM-A solution (1.2 M MgSO_4 , 10 mM sodium phosphate buffer, pH 5.8, supplemented by 25 mg/mL Lysing Enzymes from *Trichoderma harzianum*) and shaken at 100 RPM for 2 h, at 30°C. Protoplasting was confirmed by microscopy. 10 mL of Trap buffer (0.6 M sorbitol, 100 mM Tris-HCl, pH 7) was added gently without mixing before centrifugation at 4000 g, 4°C, for 15 min. The gradient obtained consists of a superior phase with the Trap buffer, an interphase, an inferior phase with the OM-A solution and a pellet with the mycelia debris. The protoplasts are present mainly in the interphase, but also in the inferior phase. The 2 phases containing the protoplasts were collected and the Trap buffer discarded. The pellet was resuspended in 10 mL OM-A solution, and the procedure was repeated one more time. One volume of STC buffer (1.2 M sorbitol, 10 mM Tris-HCl, 10 mM CaCl_2 , pH 7.5) was added to the protoplasts suspension, which was centrifuged at 6000 g, 4°C, for 10 min. Pellet was resuspended in 10 mL of STC buffer and centrifuged once again at 6000 g, 4°C, for 10 min. The pellet was resuspended in 1 mL of STC buffer. For the transformation 10 μl of an aurin-tricarboxylic acid 110 mM, 10 μg of DNA and

1×10^7 protoplasts were mixed, then incubated for 30 min at RT. 1350 μ l of PEG 60% solution was added and the tubes were incubated for 20 min at RT. Protoplasts were recovered with 2 mL of regeneration media (yeast extract 1 g/L, casamino acids 1 g/L, saccharose 342 g/L) for 18 h at 30°C and 150 RPM. Cells were plated on minimal media plates. Transformations produced between 20 and 25 recombinants each, on medium lacking uridine, to select for presence of *pyrG* following plasmid integration. For purification, transformants were propagated twice successively on the same medium. To release selective pressure on the integrated plasmid, transformants were propagated twice on Potato Dextrose Agar (PDA) containing uridine (10 mM). To select for clones that had excised the plasmid, conidia resuspended on 0.01% Tween 80 were spread on PDA containing 10 mM uridine and 1.25 mg/mL 5-fluoro-orotic acid.

A screening to find the deletion mutants was done for each transformation. The clones obtained were inoculated in TvMM liquid cultures. The mycelia were used for genomic DNA extraction, with either the classic or the rapid gDNA extraction protocol described above, and a PCR to confirm the deletion of the locus was done. A sequencing of the deleted locus was done over the selected strains to confirm the correct deletion.

2.1.6. *Hydrolytic activity test by congo red dye*

T. versatilis conidia were inoculated in TvMM supplemented with glucose 1% (w/v) and incubated for 48 h at 30°C and 150 RPM. The mycelium was then washed and transferred to TvMM with wheat straw 1% (w/v) as the carbon source as it was described before. After 24 h of incubation in the same conditions, the mycelium was filtered and the supernatant was recovered. A sample of 10 μ L of supernatant was deposited in the centre of a CMC agar (CMC 0.5% w/v, agar 1% w/v) or Xylan agar (xylan 0.5% w/v, agar 1% w/v) plate. The plates were incubated at 37°C for 2 h. They were washed with 5 mL of SSC solution (NaCl 1 M, sodium citrate 0.1 M) and 3 mL of a 5 mg/mL congo red solution were added to each plate. The plates were then incubated for 15 minutes in a shaker at room temperature. The stain was discarded and the plates were washed with SSC solution once

again. 5 mL of SSC solution were poured into each plate and they were placed in a shaker and incubated at room temperature for 15 min to allow the de-staining to occur. Finally, the hydrolysis halos were analysed and the plates were photographed.

The same test was done over the deletion mutants *T. versatilis* $\Delta xlnR$ and *T. versatilis* $\Delta clrA$. The two strains were inoculated in 100 mL of TvMM supplemented with glucose 1% and uridine 10 mM and incubated at 28°C and 150 RPM for 96 h before washing and transferring the mycelia to the TvMM with wheat straw 1%. The rest of the protocol was performed as described previously.

2.2. Basic molecular biology techniques

2.2.1. Nucleic acids isolation

2.2.1.1. Genomic DNA extraction from *Talaromyces versatilis*

Approximately 250 μ L of mycelia, collected from liquid cultures, were added to a 1.5 mL tube containing 625 μ m glass beads (Sigma). The samples were flash frozen under liquid nitrogen. To each sample, 600 μ L fungal DNA extraction buffer (Tris-HCl (pH 7.5) 50 mM, EDTA 10 mM, NaCl 50 mM, SDS 1% (w/v)) were added. The frozen mycelia samples were broken down and mixed with the glass beads and the buffer using the reverse end of a loop. The mixture was vortexed for 30 seconds twice. The RNase A (Euromedex) was added to a final concentration of 100 μ g/mL, and the tubes were incubated at 37°C for an hour. Then 100 μ L of a 4 mg/mL solution of Proteinase K (Thermo Scientific) were added to each sample and they were incubated for 3 to 4 hours at 55°C. Another incubation at 65°C for 30 minutes followed and finally, the samples were centrifuged for 5 minutes at 16000 g. The supernatant was placed in a Phase Lock tube (5 Prime) and 500 μ L of phenol:chloroform:isoamyl alcohol (25:24:1 v/v/v) were added. The tubes contents were mixed and centrifuged for 10 minutes at 16000 g. The previous step was repeated one more time. The upper aqueous phase was recovered and transferred to a new tube and 1 volume of chloroform was added. The tubes were inverted several times to mix and then centrifuged for 10 minutes at 16000 g. The aqueous phase was

removed and placed in a new tube where 0.7 volumes of isopropanol were added, the tubes were inverted several times and incubated for 30 minutes at -20°C. The samples were centrifuged for 10 minutes at 16000 g, the supernatant was discarded and the pellet was washed with 700 µL of 70% ethanol. The tubes were centrifuged one more time at 16000 g for 10 minutes and the ethanol was removed. The DNA samples were left to dry and then resuspended in 100 µL of nuclease free water. The DNA concentration and quality was assessed with the NanoDrop ND-1000 spectrophotometer (Thermo Scientific). The samples were stored at -20°C.

2.2.1.2. *Rapid gDNA extraction*

A method for rapid gDNA extraction from mycelia samples was used for the screening of strains. Mycelia samples were collected from liquid cultures or mycelia grown in PDA plates were collected using a loop and deposited in a 1.5 mL tube containing 100 µL of TE buffer. The tubes containing the mycelia samples were incubated in a water bath at 95°C for 5 minutes. They were then placed on ice and then centrifuged at 16000 g for 5 minutes at 4°C. The supernatant containing the gDNA was recovered and stored at -20°C. 2 µL or 4 µL of the gDNA samples were used for the PCR reactions in a final volume of 25 µL.

2.2.1.3. *RNA extraction from mycelia*

The frozen mycelia samples were dismembered by high-speed shaking the tubes, containing a single stainless steel bead each, for 3 minutes at 30 Hz in the TissueLyser II (Qiagen), twice. Total RNA was isolated from frozen mycelia samples using the GeneJET Plant RNA Purification Mini Kit (Thermo Scientific). An on-column DNase I treatment (Thermo Scientific) was added to the protocol. 100 µL of the DNase I mix (50 µL of DNase I, 10 µL of 10 X buffer and 40 µL of nuclease-free water) was added to the column after the first wash. The tubes were incubated for 30 minutes at room temperature and washed again with the wash buffer I. The rest of the protocol was performed as recommended by the supplier. The quantity and quality of the RNA samples were assessed by using the NanoDrop

ND-1000 spectrophotometer (Thermo Scientific) and the Bioanalyzer 2100 with the RNA 6000 Nano LabChip kit (Agilent).

2.2.1.4. *RNA extraction from conidia and germinating conidia*

For conidia and germinating conidia samples, total RNA was isolated using the TriZol reagent (Invitrogen) as follows. A total of 10^8 conidia incubated in liquid cultures for 0 h, 2 h, 4 h, 8 h, 12 h and 16 h were harvested by centrifugation at 3000 g for 2 minutes. For each sample, the supernatant was removed and the spores were mixed with 500 μ L of RNA extraction buffer (NaCl 0.6 M, sodium acetate 0.2 M, EDTA 0.1 M, SDS 4% w/v). The mix was placed in a clean 2 mL safe-lock tube and flash frozen under liquid nitrogen. The samples were slowly thawed and 625 μ m glass beads (Sigma) were added to each tube. The conidia were dismembered by high-speed shaking the tubes for 3 minutes at 30 Hz in the TissueLyser II (Qiagen), twice. The liquid phase was transferred to a tube containing 1 mL of TRizol reagent (Invitrogen) and 0.25 mL of chloroform were added to each sample. The tubes were incubated for 5 minutes at room temperature and then vortexed. The tubes were centrifuged at 16000 g for 15 minutes. The aqueous phase was transferred to a clean tube (approximately 750 μ L of sample) and 1 volume of isopropanol was added. The samples were mixed by inverting the tubes several times. The tubes were incubated at room temperature for 10 minutes and centrifuged at 16000 g for 10 minutes. The supernatant was removed and the pellet was washed with 1 mL of 70% ethanol and centrifuged once again at 16000 g for 10 minutes. The ethanol was discarded and the pellet was left to dry. Each pellet was resuspended in 50 μ L of nuclease-free water. A clean-up protocol using the RNeasy Mini Kit (Qiagen) was done following the suppliers recommendations and an on-column DNase I treatment (Thermo Scientific) was further performed on these RNA samples as described in the previous section. The quantity and quality of the RNA samples were assessed by using the NanoDrop ND-1000 spectrophotometer (Thermo Scientific) and the Bioanalyzer 2100 with the RNA 6000 Nano LabChip kit (Agilent).

2.2.2. Polymerase Chain Reaction (PCR)

All PCR reactions were run for 30 cycles in thermal cyclers (BioRad). The following programmes were used when applying the Phusion DNA Polymerases (Thermo Scientific): initial denaturation at 95°C for 5 min, denaturation at 95°C for 20 s, annealing at 50°C to 60°C, depending on the primers used, for 20 s and extension at 72°C for 1 min to 3 min, depending of the length of the amplicon. A final extension was continued for another 5 min at 72°C after cycling had finished. The PCR reactions total volume was 25 µL or 50 µL: 1 X Reaction Buffer (Thermo Scientific), 200 µM of dNTPs, 0.2 µM per primer, between 10 ng and 200 ng of template DNA and 0.05 U/µL of Phusion DNA Polymerase (Thermo Scientific).

The primers used in this work were designed using Vector NTI Advance 11 (Invitrogen) and ordered from Eurogentec (France). In general, primers were designed with the following parameters: the CG content must be 40 to 60%, with an homogenous distribution of bases, the length of the complementary region to the template should be of 18 – 25 nucleotides, there should not be any repeated sequences or auto-complementary sequences of more than 3 bases to avoid the formation of hairpin structures, the intra- or inter-primer homology has to be avoided to bypass the formation of dimers, the 3' base should be a C or a G and the melting temperatures (T_m) of both primers should not differ in more than 5°C. The primers used in this work are listed in Table 2.1.

Table 2.1: List of primers used during this work for DNA amplification

Name	Sequence	Melting T	Function
acel-fw	Fw:GAGGCTGTGTTGCGAATG	49.4	Amplification of the <i>acel</i> locus
acel-rv	Rv:TGTTGCGAAGAGTGGTGC	49.8	
pyrG-up-fw	Fw:TGTTTATCTGGCTGACATCGGCTC	52.3	Amplification of the upstream
pyrG-up-rv	Rv:GGTTCTGCAGAGGAGGACGACATTTTG	56.2	fragment of the <i>pyrG</i> locus
pyrG-dw-fw	Fw:TATCTGCAGCAGGGGTGGAAAGCCTATC	54.4	Amplification of the
pyrG-dw-rv	Rv:CGTCATCGGGTCCTTCAACGCTTC	55.7	downstream fragment of the <i>pyrG</i> locus
pyrG-fw	Fw: TGGTTTGAGACTGGCATC	53.7	Amplification of the <i>pyrG</i> locus
pyrG-rv	Rv: AGGACGACAAGGATCAGC	56.0	

xlnR-up-fw	Fw:TTCTCTGCACTAGTCGATTCCTTCTGG	54.6	Amplification of the upstream fragment of the <i>xlnR</i> locus
xlnR-up-rv	Rv:TTATAGAAGGGCCCAACCGAAGTCCAC	56.2	
xlnR-dw-fw	Fw:AACGACGGGCCAGGTTCTTGCACTTTATC	59.3	Amplification of the downstream fragment of the <i>xlnR</i> locus
xlnR-dw-rv	Rv:AACCTCCACTAGTCAGCGTCTTCAGCCG	59.2	
xlnR-fw	Fw:CTGTGATTCGGTTGCTTTCTGG	49.7	Amplification of the <i>xlnR</i> locus
xlnR-rv	Rv:TCGCATCGCTAACAGACAACG	49.2	
clrA-up-fw	Fw:GTACCGGGCCCCCTCTGTATAATTTTGCTCC	57.5	Amplification of the upstream fragment of the <i>clrA</i> locus
clrA-up-rv	Rv:CCGTCGAAGAAAATCCAGCCAGGGG	63.1	
clrA-dw-fw	Fw:GATTTTCTTCGACGGACATTGCGAGG	54.4	Amplification of the downstream fragment of the <i>clrA</i> locus
clrA-dw-rv	Rv:CGGGCTGCAGGAATTAGGTTGATGTCTGAAGC GAAG	62.8	
clrA-fw	Fw:TGATTGCGTCTTGGGAGC	56.0	Amplification of the <i>clrA</i> locus
clrA-rv	Rv:GGTTGACTTGACTGGCTATGG	59.8	
creA-up-fw	Fw:CTTGATATCGAATTCGCTCTTACAGGGCT CAAGTC	61.6	Amplification of the upstream fragment of the <i>creA</i> locus
creA-up-rv	Rv:TCTGTCACTGCTCCCACCGACTCGCTCGGCT CCAGTTTC	68.4	
creA-dw-fw	Fw:GAAACTGGAGCCGAGCGAGTCGGTGGGAGC AGACTGACAGA	68.4	Amplification of the downstream fragment of the <i>creA</i> locus
creA-dw-rv	Rv:GGGCTGCAGGAATTCCTCCGCTCAAAGTATCGT CATCATG	64.7	
creA-fw	Fw:AAATCCCACACCATCCCG	56.0	Amplification of the <i>creA</i> locus
creA-rv	Rv:CCTTCCATCTATCTCCCTACGG	62.1	

2.2.3. *Plasmid construction*

The different plasmids used in this work were constructed by modifying the pC3 plasmid (Delmas *et al.*, 2014) designed for the genetic modification of filamentous fungi. A list of the plasmids used is details in the Table 2.2.

For the plasmids construction, the PCR amplified fragments were cloned into the pC3 vector using either standard procedures (Sambrook & Russell, 2001) or the In-Fusion HD cloning kit (Clontech) following the manufacturer's recommendations.

E. coli XL10-Gold competent cells (Stratagene) were transformed with the plasmid constructed by classic procedures. Stellar competent cells (Clontech) were transformed with the plasmids constructed by the In-Fusion method, following the recommendations of the supplier. In both cases, the transformants were selected in LB plates supplemented with ampicillin 150 µg/mL, x-gal 40 µg/mL and IPTG 0.2 mM. White colonies were selected and inoculated into 5 mL of LB broth containing 150 µg/mL of ampicillin. The cultures were incubated at 37°C for at least 16 h, with shaking at 200 RPM. Plasmid DNA was then purified using the GeneJET Plasmid Miniprep kit (Thermo Scientific), following the supplier's recommendations. The plasmids were stored at -20°C. Restriction digests (New England Biolabs) were carried out following the manufacturer's instructions to confirm the insertion of the fragments in the right position. The plasmid DNA band patterns were visualized in agarose gels 0.8% (w:v).

Table 2.2: Plasmids used for the genetic modification of *T. versatilis*

Plasmid	Comment	Insertion	Function	Source
pC3	Derived from the pBluescript SK ⁻	<i>A. oryzae pyrG</i> locus		(Delmas <i>et al.</i> , 2014)
pC3-Tv_pyrG	Derived from the pC3	<i>pyrG</i> cassette with the up- and downstream regions of the <i>T. versatilis pyrG</i> locus	Integration into the genome and complementation of the <i>pyrG</i> deletion	This work
pC3-Tv_xlnR	Derived from the pC3	<i>xlnR</i> cassette with the up- and downstream regions of the <i>T. versatilis xlnR</i> locus	Integration into the genome and deletion of the <i>xlnR</i> locus	This work
pC3-Tv_clrA	Derived from the pC3	<i>clrA</i> cassette with the up- and downstream regions of the <i>T. versatilis clrA</i> locus	Integration into the genome and deletion of the <i>clrA</i> locus	This work
pC3-Tv_creA	Derived from the pC3	<i>creA</i> cassette with the up- and downstream regions of the <i>T. versatilis creA</i> zinc-fingers domain	Integration into the genome and deletion of the <i>creA</i> zinc-fingers domain	This work

2.3. Gene expression analysis

2.3.1. PCR based methods

2.3.1.1. Primer design for mRNA quantification

For RNA quantification, when it was possible, the primers were designed in such a way to minimize the amplification of the contaminating genomic DNA. To do this one of the primers in each pair was designed in an exon-exon junction. All the primers were designed following the parameters described previously. For each gene of interest 2 to 3 pairs of primers were designed, their efficiency and quality was tested. The best primers, that produced no primer-dimers and that were not sensitive to the gDNA, were chosen for further works. A list of all the genes studied during this project and the primers used is given in Table 2.2.

Table 2.2: Primers used for mRNA quantification

Name	Gene	Sequence	Product length	Exon-exon junction
R1	DUF221 domain protein (<i>DUF221</i>)	Fw: CGGAACGCCCCATTGACC Rv: TTGGATGCTTATGTTTTGCTCTCG	126 bp	Yes
R2	Ubiquitin carrier protein (<i>ubcB</i>)	Fw: TCGTTGAGTAGACTCTGAATGCTG Rv: AGCCAGATGTTCCACCCG	125 bp	Yes
R3	CECR1 family adenosine deaminase (<i>ADA</i>)	Fw: CTGCGCAATGCAAAGTCATGTCTCTG Rv: CCCAGGTGCAAGATCCCTTATCCA	97 bp	Yes
R4	Mitochondrial membrane fission protein (<i>fis1</i>)	Fw: GTTCAACTACGCCTGGGGACTC Rv: AGCGGTGCGAAAAATCTGGG	91 bp	Yes
R5	Copper-transporting ATPase (<i>Cu-ATPase</i>)	Fw: TGGTGCCTGTGCCAACTCTCCCAGTC Rv: TTGCTGCGGGTGCTTTTG	78 bp	Yes
R6	Cohesin complex subunit (<i>psm1</i>)	Fw: GTATTTGCGGAGATCCAGAGTGAG Rv: TTGAAGACGGGTCTGTTCCA	102 bp	Yes
R7	Spo7-like protein (<i>spo7</i>)	Fw: GCCGATGGTGCTGATGTTGG Rv: AGAACGCCAACGAGCCCG	110 bp	No
R8	SAGA-like transcriptional regulatory complex subunit Spt3 (<i>spt3</i>)	Fw: ACGACTTGTTGGCGGACG Rv: GAGATTCAGCAGATGATGTTTGTC	95 bp	Yes
R9	DUF500 domain protein (<i>DUF500</i>)	Fw: ACTTGCCGGTTGTGCGTTC Rv: TTGGTGTTCCGGCGGCTG	101 bp	No

R10	Rho GTPase activator (<i>sac7</i>)	Fw: AGGAGGATGAAAGTAAAGGACCCC Rv: AAACCCACACTTGGCGAC	159 bp	Yes
R11	AP-2 adaptor complex subunit beta (<i>AP-2 β</i>)	Fw: TTTCGCACATAGGGGTCG Rv: TTTTGGTTCGATGATATGGACG	148 bp	Yes
R12	Protein translocation complex componenet (<i>np1</i>)	Fw: CGCTGGAACAAGAAAAATACG Rv: ACGAACGATATGCGCCAA	117 bp	Yes
Tub	β-tubulin	Fw: GTTCTGGACGTTGCGCATCTG Rv: TGATGGCCGCTTCTGACTTCC	110 bp	Yes
GOI2	<i>cbhl</i> – GH6	Fw: TCGTGGTTGTTGTTGGCG Rv: CAGCACTCAAACCCCTTACTACGCAC	119 bp	Yes
GOI4	<i>cbhl</i> – GH7	Fw: GGTGTATCGCAAGGGTGAGGAGTC Rv: ACGGATCTTGCTGCGCGG	82 bp	No
GOI6	<i>abf-B2</i> – GH54	Fw: CGGAGCTTGGGTGAGATGGTTC Rv: CGGCGGCGTTGCTAATGC	112 bp	No
GOI7	<i>abf-B1</i> – GH54	Fw: ACCACCCGTGCCCTTTATGC Rv: CGGAAAGCGGCGAGATGG	94 bp	No
GOI10	<i>xynB</i> – GH11	Fw: ATTCACGAGCGGGTTGG Rv: ACCAGGTGATGCTAACCCCATC	101 bo	Yes
GOI12	<i>xynC</i> – GH11	Fw: CAAATGGCGACAATGGCG Rv: TGAGTACGTGACAGTCTGTGCATTG	104 bp	Yes
GOI14	<i>xynD</i> – GH10	Fw: ATTGCCGTCATACACTTTCG Rv: CGTCTTCGTTCAAGGCTTCGTTTAC	143 bo	Yes
GOI15	β-glucosidase – GH1	Fw: GGTGTGCATGAGCCGTGG Rv: TCGGCCAGTCACCATTGAG	134 bp	Yes
GOI19	β-xylosidase – GH3	Fw: TTTGCACGGGCGATAGGG Rv: CTGTGGAGGGACCTGGCTAGTTG	156 bp	Yes
GOI23	β-xylosidase – GH43	Fw: TGCCCCGAAGATTGAAGGACC Rv: CCAGCGAATAAACGGGGTTCG	147 bp	No
GOI26	Endoarabinase – GH43	Fw: CATTGCTGCGGATACGACTTC Rv: CATTGGCGTCAACGAAACCG	104 bp	Yes
GOI27	α-L-arabinofuranosidase - GH43	Fw: ACCATGACAACACCGGAGCTACC Rv: TGCCAATTCGCCATATCTGC	79 bp	No
GOI30	GH61	Fw: AAGTCGCTGCCCTTTCTGCC Rv: CAAGGTATCCAGAGTAGGATTCACCG	105 bp	Yes
GOI31	Hydrophobin	Fw: CCGCGACTTTGGATGACTGAGC Rv: CCGCCTTTGGTGACGCTGAC	111 bp	Yes
GOI33	Swollenin	Fw: CCGCTCCAGATCCACAACACC Rv: TATGCAGTGGGCGAGGATGG	103 bp	Yes
VPA	Intergenic region	Fw: ACCGAATGGCACCGAGTTGG Rv: AATGGAGGAAGCGTGCCGTG	133 bp	--

2.3.1.2. Reverse-transcription

For cDNA synthesis only the RNA samples with 260/280 nm wavelength ratio of approximately 2 and 260/230 nm wavelength ratio greater than 2 were used. Synthesis of cDNA was performed using

the PrimeScript First Strand cDNA Synthesis Kit (Takara), following the manufacturer's protocol. One microgram of total RNA from mycelia samples and 100 ng of total RNA from conidia and germinating conidia were used for the cDNA synthesis reaction. The cDNA was diluted 1:10 with water and stored at -20°C.

2.3.1.2. Semi-quantitative PCR

A classic PCR reaction was done over the cDNA samples, as described previously. In this case, 1 µL of cDNA was used as template in a 25 µL PCR reaction. For the quantification of each gene of interest, the PCR reaction was stopped after 15, 20, 25 and 30 cycles. The amplicons were visualized in 0.8% agarose gels. The intensity of the bands was compared between the different cycles to assess the relative amount of cDNA present in the sample; if the bands were visible at an early cycle it indicated that the amount of cDNA was higher.

2.3.1.3. Quantitative PCR

The qPCR was performed in a CFX96 Real Time PCR Detection System (Bio-Rad), using 96-well white PCR plate (Thermo Scientific) sealed with Absolute qPCR seals (Thermo Scientific). The reaction mix consisted of 7.5 µL of the DyNamo ColorFlash SYBR Green master mix (Thermo Scientific), 300 nM of each primer and 3 µL of the 1:10 diluted cDNA in a final volume of 15 µL. The PCR reaction cycle was: initial denaturation for 7 minutes at 95°C, followed by 40 cycles of 10 seconds at 95°C and 30 seconds at 60°C. A melting curve was performed at the end of the qPCR run, increasing the temperature in a stepwise fashion by 0.5°C every 5 seconds, from 65°C to 95°C. Each qPCR reaction was performed in technical triplicate. Two control samples were included for each primer pair tested: the no template control (NTC) and *T. versatilis* genomic DNA.

For each sample, a ValidPrime Assay (VPA), consisting of a pair of primers that bind to a non-transcribed intergenic region identified from RNA-seq data, was also included to detect and

quantitate the presence of contaminating gDNA (Laurell *et al.*, 2012). As gDNA contamination rarely exceeded 1%, the qPCR data were directly analysed using the CFX Manager software (Bio-Rad).

2.3.2. RNA sequencing method

2.3.2.1. Experimental design

T. versatilis conidia were inoculated at a final concentration of 2×10^5 spores/mL in 3 x 100 mL TvMM complemented with glucose 1% (w/v) and incubated for 48 h at 28°C and 150 RPM. Glucose-grown mycelia from two flasks were rinsed with 100 mL TvMM without any carbon source and transferred into 2 x 100 mL TvMM supplemented with wheat straw 1% (w/v). The flasks were incubated for 24 h at 28°C and 150 RPM. Then 1% (w/v) glucose was added to one of the flasks and it was incubated for 5 h at 28°C and 150 RPM. After each step, mycelia were harvested, snap frozen in liquid nitrogen and stored at -80°C. Each growth step was carried out in triplicate. RNA was extracted as described before from mycelia grown in triplicate from each of the three conditions described above.

2.3.2.2. Sequencing

SOLiD whole transcriptome libraries were made according to the SOLiD Whole transcriptome kit protocol (Applied Biosystems) and the library concentrations were measured with the Quant-ut HS dsDNA assay kit (Invitrogen). Libraries were pooled to equimolar amounts (Invitrogen) and gel purified using 2% size-select E-gels to 200-300 bp (Invitrogen). Emulsion PCR (0.5 M final concentration of pooled libraries) and bead-based enrichment was done according to the SOLiD 3+ template bead preparation guide containing library. Sequencing was carried out on a SOLiD 3+ ABi sequencer platform according to manufacturer's instructions to generate 50 bp reads in colour space. The sequencing and the bioinformatics analysis that followed was performed by the University of Nottingham Deep-Seq facility.

2.3.2.3. First analysis of the sequencing results

RNA-seq reads from SOLiD from each of the experimental samples were mapped to the *T. versatilis* genome using the BioScope 1.3.1 Transcriptome Pipeline (LifeTechnologies). Bioscope mapped each read against the complete genome sequence and exon spanning junctions using available gene coordinate information, hence providing the primary read alignment position. Mapping results were recorded in a BAM (binary alignment/map) format.

The amount of reads per gene for each sample was calculated with Htseq-count (<http://www.huber.embl.de/users/anders/HTSeq>) using the BAM files and the genome annotation. Strand-specific RNA-seq reads, as generated by SOLiD, can be specified for when executing Htseq-count in determining accurate read counts per gene. These counts were used to calculate normalised expression values (reads per kilobase of exon model per million mapped reads = RPKM) for each gene (Mortazavi *et al.*, 2008). Antisense transcription was detected by either excluding or including strand-specificity in the calculations when comparing the Htseq-count generated counts. Three independent statistical tests were carried on the number of reads counted per gene for each condition. These tests were the Likelihood Ratio Test (Marioni *et al.*, 2008), Fishers's Exact Test (Bloom *et al.*, 2009) and an MA-plot-based method with the Random Sampling method (Wang *et al.*, 2010b). These tests allowed to find all the genes which are significantly differently expressed for all three conditions studied and all those genes scored a p-value of >0.001.

Files containing the RPKM value per gene and the annotated genome were used to map gene I.D., annotation and expression value (using the average from triplicates) to each other using the Excel "lookup and reference" functions.

2.3.2.4. Annotation pipeline

SOLiD reads were first processed using the SOLiD Accuracy Enhancement Tool (SAET). Reads were then filtered against a collection of fungal rRNA and tRNA sequences as well as sequencing library adaptors. Reads were then filtered according to colour call quality using stringent criteria. This

yielded 171,477,828 of the total 368,065,069 reads to be used in defining the transcriptome. Filtered SOLiD reads were aligned to the genome sequence using the RNA-seq alignment tool TopHat (v2.0.7). The resulting alignment BAM file was filtered to remove low quality alignments using MAPQ20 as a threshold. The tool Cufflinks (v.2.1.1) was used to process the filtered BAM file into a number of transcripts. De Novo Gene predictions were made using the package GeneMark.

2.3.2.5. *Second analysis of the RNA-seq data*

With the new *T. versatilis* genome annotation, a second analysis of the sequencing results was performed. The LifeScope (v2.5.1) Whole Transcriptome Pipeline (Life Technologies) was used to filter and then map the reads onto the reference genome, generate read counts per gene and RPKM expression values. Reads were first filtered against a set of sequences to remove those reads derived from undesired sequences, such as rRNA, tRNA and sequencing library contaminants. Reads that pass the filter were mapped against the reference genome in the context of known gene exon coordinates. Reads are also aligned to a library of exon junctions derived from the exon coordinates in the GTF file, thereby allowing reads that map across exon boundaries (spliced reads) to be recorded. The read alignments are recorded in a BAM formatted alignment file. BAM files can be viewed visually when accompanied with the reference genome sequence, using the IGV software (<http://www.broadinstitute.org/igv/home>). Read counts for each gene are calculated using 'htseq-count'. The program determines the number of uniquely aligned reads per gene.

Read alignments, both primary and unique, have been filtered further according to their mapping quality score (MAPQ). MAPQ20 was used in our data, which indicates that there is 1 in 100 chance of a mapped read being incorrectly placed.

2.3.2.6. *Differential gene expression*

Significantly differentially expressed (DE) genes were calculated using the R Bioconductor package DESeq version 1.9. The uniquely aligned read counts per gene (MAPQ20) for each sample

replicate were used as the input to the analysis. An adjusted threshold p-value (p_{adj}) of 0.05 was used to determine significantly differentially expressed genes. This list of genes has also been sorted according to up and down changes in gene expression.

2.3.2.7. GO enrichment analysis

Enrichment of significantly differentially expressed genes in terms of Gene Ontology (GO) database terms was determined using the R bioconductor package GOseq. GO categories either significantly enriched or depleted in each sample group comparison was determined using a p-value of 0.05.

2.4. Computational analysis

2.4.1. Statistical analysis

Statistical analyses were conducted by using the STATGRAPHICS Centurion 16 software. This included the statistical analysis in chapter 5: the ANOVA on relative FC values (Three-level, nested ANOVA with 'genes' as the first level, 'culture conditions' as the second level and 'biological replicates' as the third level); the Hierarchical Ascendant Classification (HAC) that was performed according to the Ward method, using default parameters (standardization of the data and squared Euclidean distances); the graphical representation of box plots. The statistical analyses in Chapter 6 were performed using the R software, thanks to the collaboration of Sébastien Déjean (Biostatistics Genotoul Platform, University of Toulouse). The RT-qPCR data, presented in chapters 6 and 7, were directly analysed using the CFX Manager software (Bio-Rad), which allows inter-run calibrations, efficiency correction, normalization with multiple reference genes and calculation of ratios with (technical) errors propagation.

2.4.2. Genomic sequence analysis

The *T. versatilis* genomic sequence was available in the Biomax Informatic (Germany) platform. Sequences of the genes of interest were imported and analysed using the Vector NTI Advance 11 software (Invitrogen). Different bioinformatics tools were used to determine if the annotation of the genes was correct: BLAST, multiple alignments, analysis of the Pfam domains and GO terms. When it was necessary, we introduced corrections to the annotation: changes in the start and stop codon positions, changes in the reading frame, and changes in the positions of the intron-exon boundaries. Since no changes could be directly introduced in the Biomax platform, the corrected annotations were stored in a Vector NTI Advance 11 database.

3. *Talaromyces versatilis* mycelia growth and development

3.1. *Talaromyces versatilis* growth with different carbon sources and comparison to other filamentous fungi

To study the influence of different carbon sources on *T. versatilis* mycelium development, the conidia from the industrial strain were inoculated on solid minimal media supplemented with different carbon sources (Figure 3.1). A hundred conidia were deposited in the centre of each plate as described in the materials and methods section (chapter 2).

In all plates, the mycelium observed was flat, with a circular shape, and entire margins. The growth of *T. versatilis* was rather slow with all the carbon sources tested, although after 3 days of incubation at 28°C the mycelium was already visible at the inoculated spot. At this point, it was already evident that *T. versatilis* could grow much faster in the media with wheat straw. After 7 days of incubation, the plates with wheat straw and cellulose showed the best growth rate, followed by xylan, arabinose, glucose and, in the last place, xylose. The mycelium seemed to take a long time to start to grow on xylose. Similarly to what we could observe with the other 2 monosaccharides, glucose and arabinose, the mycelium on xylose was very compact and thick. Ten days after the inoculation, the plates with wheat straw and cellulose were completely covered by the mycelium, which showed a cottony appearance, in contrast to the other plates, where the mycelia were very thick. Spores could be observed after 10 days of incubation on most carbon sources as a dusty layer spreading out from the centre of the colony, except for wheat straw where no spores could be observed.

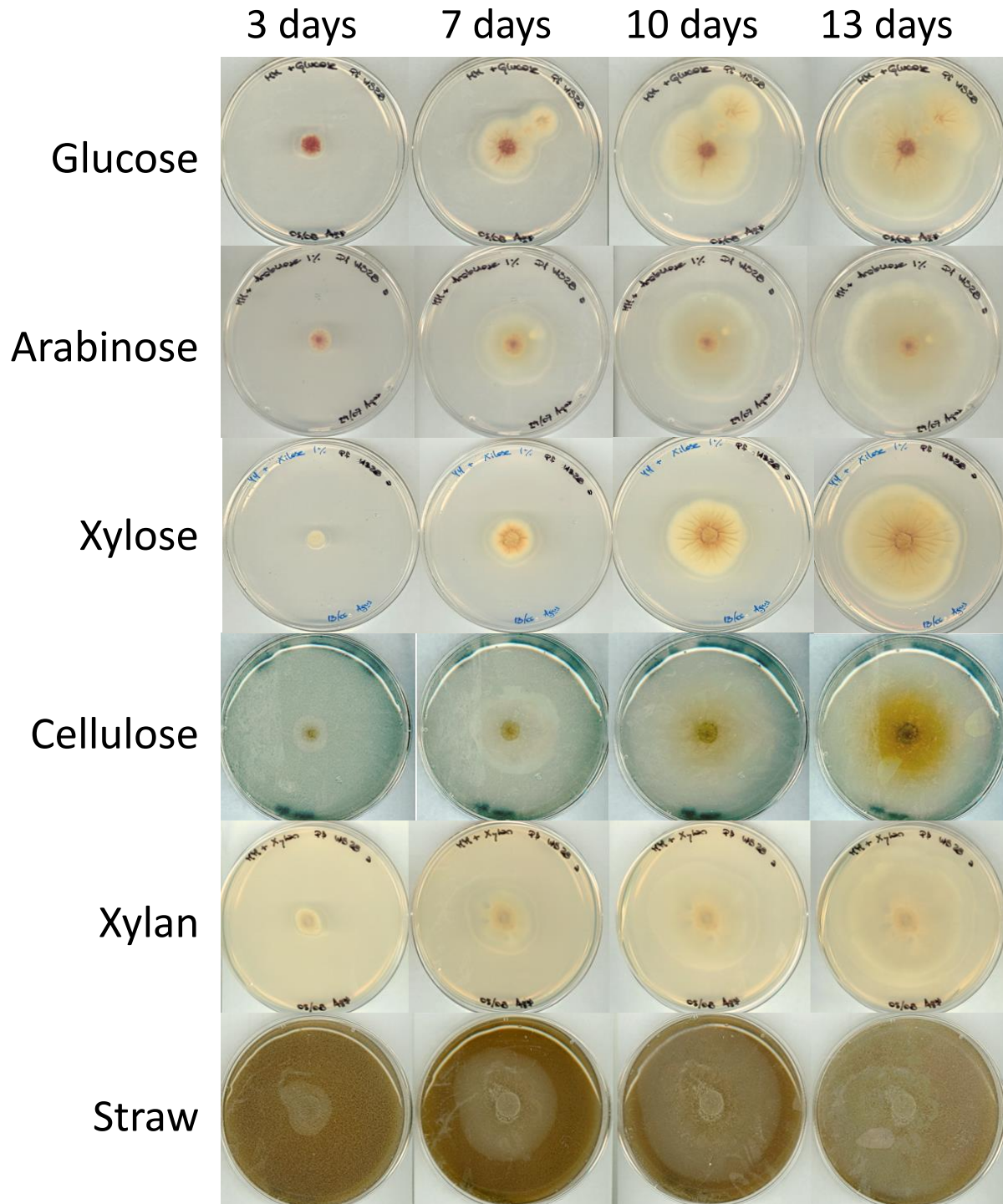


Figure 3.1: Growth of *T. versatilis* on different carbon sources. Minimal media agar plates supplemented with 1% of the different carbon sources: glucose, arabinose, xylose, cellulose, xylan and wheat straw. A hundred *T. versatilis* spores were inoculated in the centre of each plate and growth was followed at 28°C for 3; 7; 10 and 13 days.

Plates containing *Trichoderma* (TMM) or *Aspergillus* (AMM) minimal media (see chapter 2) supplemented with different carbon sources were inoculated with spores from other filamentous fungi, namely, *Trichoderma reesei* QM6a (on TMM plates), *Aspergillus niger* 8.1 and *Aspergillus nidulans* 2.137 (both on AMM plates). A hundred spores of each fungus were inoculated in the centre of the plate and they were incubated at 28°C. The growth of the three fungi on the different plates is shown in Figure 3.2.

T. reesei showed a very fast growth rate with all the carbon sources tested. After 7 days of incubation all the plates were covered by the mycelium and spores were visible on top of the mycelium. It is interesting to notice that cellulose and wheat straw plates were covered with a very fluffy mycelium, almost transparent, through which it was possible to see the agar underneath. Conversely, the plates with xylan, glucose, arabinose and xylose were covered with a cottony mycelium, but with a higher density of hyphae that did not allow to see the plate by transparency. Xylan was the substrate that induced the highest production of spores that are visible as a dark green layer on the plate and that were already present after 3 days of incubation. Spores were also observed on media with cellulose, wheat straw, arabinose, xylose and glucose, but in a lower quantity and only after 7 days.

The second fungus, *A. niger*, showed a slower growth rate on plates than *T. reesei*. In fact, this fungus did not succeed to fully cover any of the plates after 7 days. Only after 2 weeks of incubation on wheat straw, the mycelium reached the borders of the plate (not shown). The growth on arabinose and cellulose was very limited and slow. Whereas on glucose and xylose, *A. niger* produced a thick and dense mycelium. On xylan and wheat straw the growth was good, but the mycelium had a fluffy appearance. Nevertheless, black spores were produced with all the substrates. In particular with xylan and wheat straw, the spores appeared very soon, and were easily perceptible after only 3 days.

A. nidulans presented a very good growth rate that started very fast; after 3 days the colonies in all the plates already had a considerable size. After 7 days, most of the plates were almost entirely covered. Only on wheat straw the mycelium grew over the entire plate, although after 2 weeks of incubation all the other plates were also covered (not shown). Even though the growth over all the carbon sources was very good, the appearance of the mycelia was radically different. On glucose, for instance, mycelium was dense, with a pink colour and green patches of spores and the colony borders were irregular, whereas in all the other carbon sources the colony had more regular borders, though the form of the colony was never perfectly circular. In wheat straw, cellulose and arabinose the colonies presented a cottony and open mycelia. Spores appeared in all plates, but on xylan and wheat straw, the spores were present after only 3 days of incubation, as it was observed in *A. niger*. Also, xylan induced the production of a great number of spores, which were visible as a grey-green layer on top of the mycelium.

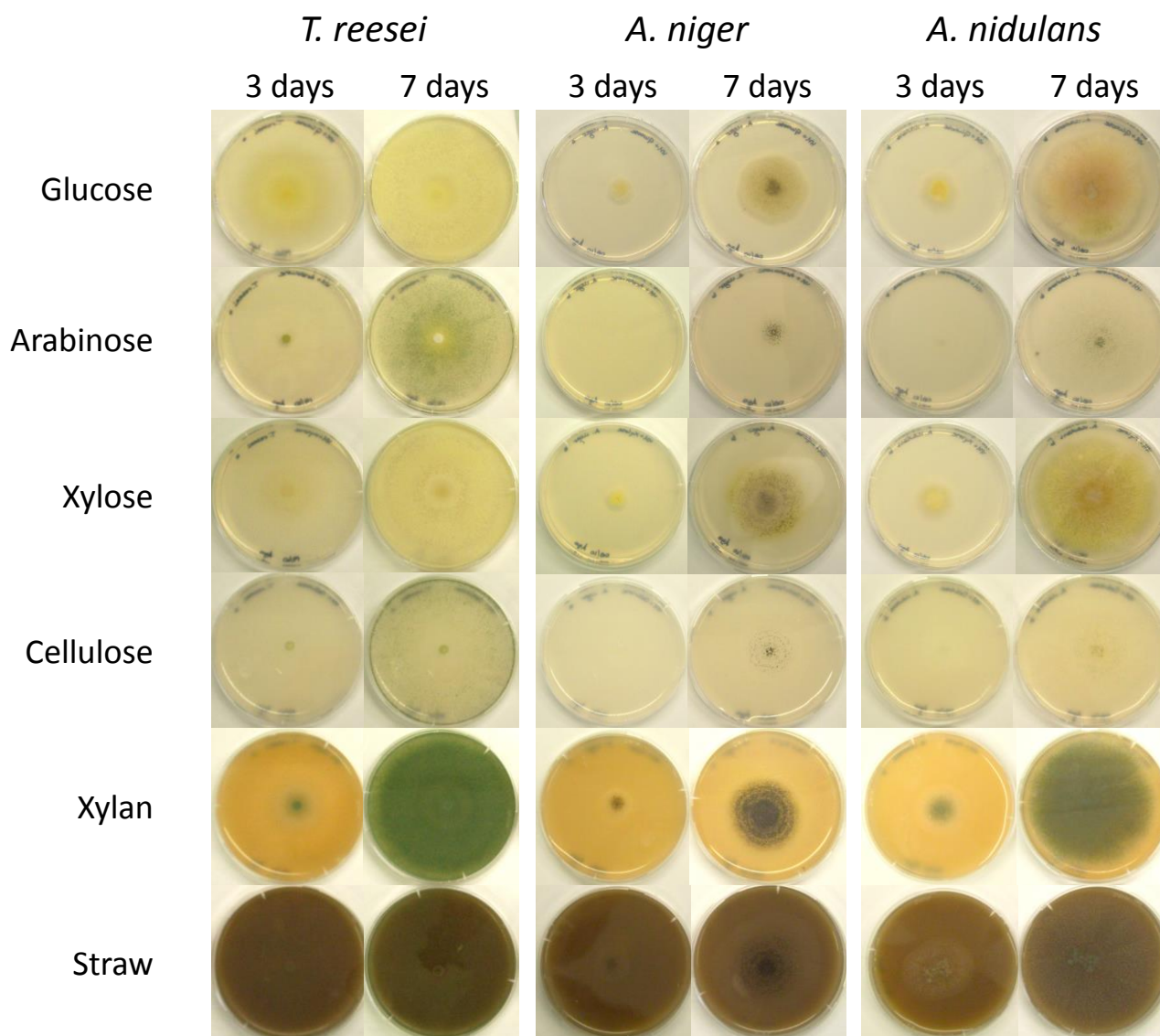


Figure 3.2: Growth of filamentous fungi on different carbon sources. Minimal media agar plates supplemented with 1% of the different carbon sources: glucose, arabinose, xylose, cellulose, xylan and wheat straw. A hundred *T. reesei* (on the left), *A. niger* (in the centre) or *A. nidulans* (on the right) spores were put in the centre of each plate. Incubation at 28°C and mycelia growth shown after 3 and 7 days of incubation.

3.2. Growth on liquid media with different N sources

The growth of *T. versatilis* on liquid media was also studied. Two different minimal media compositions were used: the *T. versatilis* specific minimal media (TvMM), which contains ammonium as nitrogen source; and the Aspergillus minimal media (AMM), which has nitrate instead. The rest of

the salts present in both media do not present major differences. Both media were inoculated with *T. versatilis* conidia to a final concentration of 2×10^5 conidia/mL. The mycelium was collected at different time-points and dried. The dry weight was plotted against the time on Figure 3.3.

The fungal biomass reached a level that could be accurately weighed after 40 hours of incubation at 28°C and 150 RPM. At this point the fungus entered the exponential growth phase. But 10 hours later the growth rate started to decline and the fungus entered in the stationary phase. At 75 hours the growth was completely stopped and the mycelium entered the autolytic phase. After 100 hours of incubation there was not enough mycelium to be accurately measured. The highest amount of biomass was detected on TvMM, and it reached 4.1 g/L (dry mass). On the *Aspergillus* minimal media, only 3.3 g/L of mycelium could be obtained. These results indicate that *T. versatilis* prefers the ammonium as N source to the nitrate. Also, the doubling time seemed to be considerably shorter on TvMM, hence the specific growth rate must have been much higher on this media. The few time-points obtained did not allow an accurate calculation of these two parameters, however an estimation of the doubling time was calculated: 10.7 h in TvMM and 16 h in AMM.

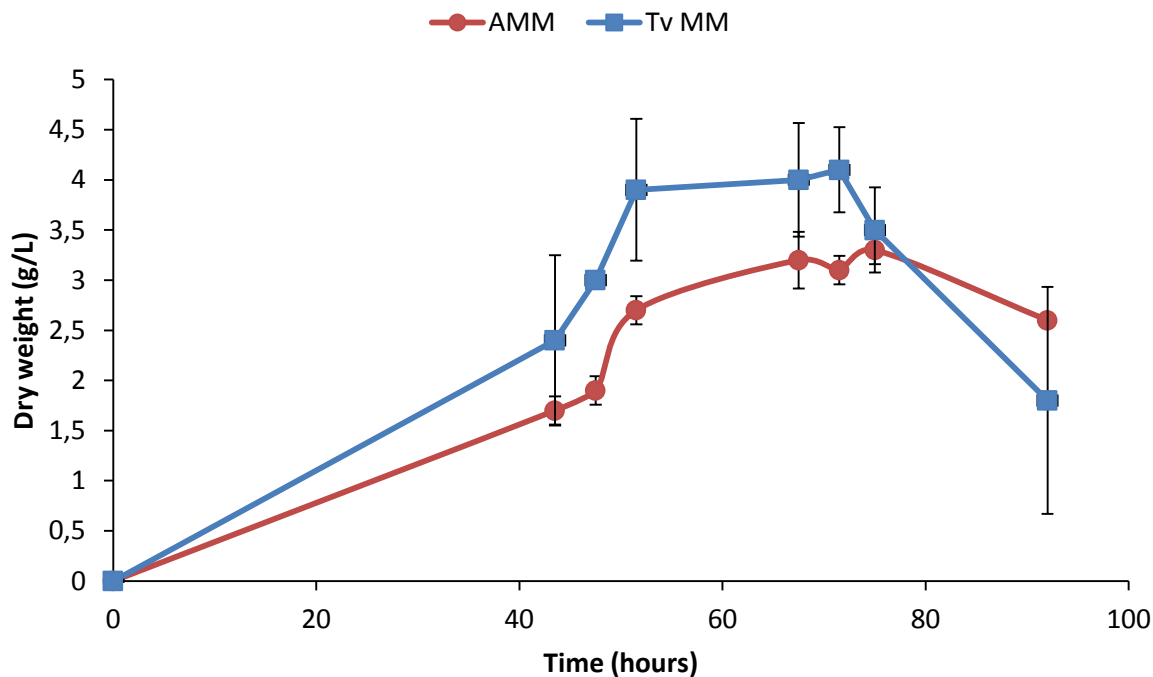


Figure 3.3: Growth curves of *T. versatilis* in TvMM and AMM supplemented with 1% glucose. Standard deviation between the 2 biological replicates is shown as vertical error bars.

3.3. Discussion

The growth of *T. versatilis* on plates was extremely slow compared to the other fungi used in this experiment, no matter which carbon source was used. Even for wheat straw and cellulose that are the best substrates for *T. versatilis*, it takes much longer (2 weeks) to the mycelia in those plates to expand and cover the whole surface of the agar. Whereas *T. reesei*, for instance, grows very fast and the plates are covered with conidia in just a week. However, when we observe the *T. versatilis* mycelia on the TvMM plates it is surprising to see that this fungus grows better on the most complex carbon sources, cellulose and wheat straw, which are usually the most difficult to degrade. This suggests that *T. versatilis* easily produces the necessary enzymes to degrade the substrate before being able to incorporate smaller saccharides into the cell for energy and biomass production. The ability of the industrial strain to grow on these conditions may be related to the fact that this strain is

the result of a mutagenesis process in which the selection of enzymes-overproducing strains was done over complex substrates.

The growth of other filamentous fungi was also studied: *A. nidulans*, which is a model fungus and *T. reesei* and *A. niger*, which are industrially important fungi, although the strains used here are not the industrial strains. All this fungi belong to the Ascomycota division, like *T. versatilis*. But only the two Aspergilli are found in the same family as *T. versatilis*: the Trichocomaceae family. Whereas *T. reesei* belongs to the Hypocreaceae family and it is rather distant from the other three fungi included in this study.

The three other fungi were also able to grow on wheat straw and cellulose, but the mycelia in those media was fluffy. The hyphae have a lot of space in between, which gives the mycelia a cottony appearance and it is even possible to see the agar surface through the colony. This might be related to the fact that the hyphae do not produce many branching points in these conditions, like if the fungus keeps extending looking for a more suitable substrate. In fact, in a media where nutrients are available the branching frequency is higher in order to fill the space and exploit the nutrients efficiently. But the frequency of branching is lower when the nutrients are scarce (Kavanagh, 2011). In the case of the cellulose and the wheat straw, the polysaccharides present need to be degraded prior to their incorporation and utilisation by the fungus, this makes them a difficult substrate to use. The fact that the *T. versatilis* mycelium is thicker in these conditions suggests that the branching frequency is higher. More branching also means more hyphae tips that can secrete proteins to the media. Clearly, *T. versatilis* is better prepared to make use of this kind of substrates, and this includes the production and secretion of hydrolytic enzymes that will allow the degradation of the polymers to produce smaller molecules, which are easily incorporated by the fungal cell.

The growth of *T. versatilis* in liquid media supplemented with the two different nitrogen sources, ammonium and nitrate, is very slow. The doubling time is extremely high in both media tested and the maximal amount of mycelium produced per litre is not very important. Nevertheless, the growth

is better on the TvMM than on the AMM culture. This is probably due to the presence of ammonium as a nitrogen source in the TvMM, which seems to be more suitable for *T. versatilis* than the nitrate present in the AMM. Every fungus is capable of using an organic source of nitrogen, but not all of them can use inorganic nitrogen. Nitrate can be utilised by numerous fungi; after its incorporation into the cell it is first converted to nitrite and then to ammonium. The ammonium can be combined to organic acids to produce amino-acids, usually glutamic or aspartic acid. The other amino-acids can be produced by transamination reactions (Deacon, 2006). Therefore, ammonium is a more convenient nitrogen source and clearly *T. versatilis* has difficulties in producing the necessary enzymes for the conversion of nitrate into ammonium. The presence of nitrate in the AMM media increases the doubling time 6 hours, from ≈ 10.7 h in the TvMM to ≈ 16 h in the AMM.

4. Genomic sequence annotation

4.1. Manual curation of the genome

In 2007 Adisseo decided to start exploring the genomic sequence of the *T. versatilis* industrial strain, therefore a first sequencing was carried out using the Roche/454 technology (Guais, 2009). Right before the beginning of the project presented here, a second sequencing was carried out using the Solexa/Illumina technology which helped to improve the quality of the genomic sequence and an automatic annotation was also done. An improvement in the annotation in order to correct the structural annotation of genes is critical to carry on the downstream functional genomics approaches (Cerqueira *et al.*, 2014). The annotation of the latest genomic sequence of *T. versatilis* presented several problems, starting with the overestimation of the number of genes. The closest filamentous fungi to *T. versatilis* whose genomes were available, the *Aspergilli* and *Penicillium*, have between 9000 and 12000 genes annotated and the genomes are about 30 - 40 Mb. *T. versatilis* has over 18000 genes automatically annotated, many of them with unknown function and very weak homology to the annotated genes in other fungi. Furthermore, in some cases the boundaries of the genes were not well established, as well as the intron-exon limits. Another inconvenient was the ORFs that were annotated in the same region, that is to say, superposed. All this errors needed to be manually curated.

In view of the vast work that this task would represent, we decided to focus on three main groups of genes that are the centre of our interest: the glycoside hydrolase (GH) genes, the transcription factors involved in the regulation of the expression of GH-encoding genes and the secretion genes. Each one of these categories will be discussed in the following sections.

Using BLAST and sequence alignments tools the genes of interest could be identified and then corrected (Figure 4.1). Each ORF, with its correct limits and the limits between the introns and exons,

was annotated using the Vector NTI 11 software (Invitrogen). Since we only had access to view and extract sequences from the annotated genome provided by Biomax Informatics (Germany), and it was not possible to modify the annotation of the genes, the correct annotation plus the upstream and downstream sequences were stored in the Vector NTI software database. The example presented in Figure 4.1 corresponds to an annotated gene that corresponded to 2 different genes. The separation of the two genes was necessary, although this leads to an increase in the gene number. This kind of situations was found in several cases, but for other loci the opposite situation was encountered: two different genes were annotated in the same locus, one of them was real and the other one not.

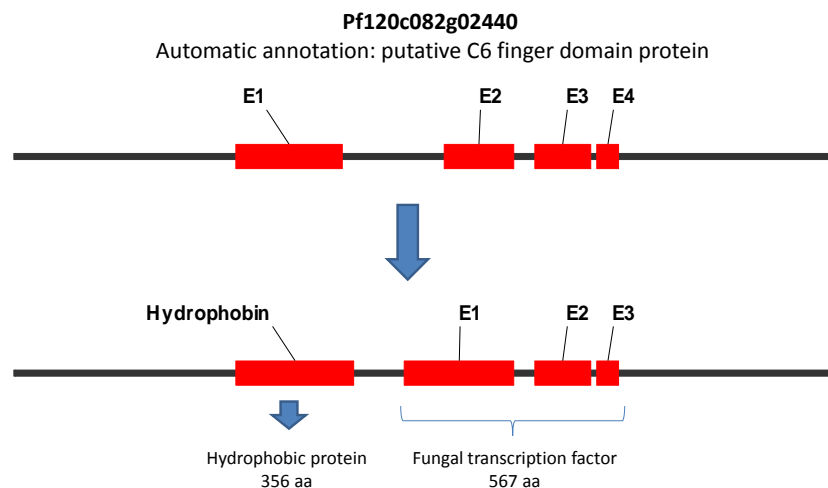


Figure 4.1: Example of the manual correction of the genes in *T. versatilis*. A gene that was annotated as a “putative C6 finger domain protein” and that actually had a “fungal transcription factor” domain and a “hydrophobin” domain, was split in two different ORFs. The stop codon of the hydrophobin encoding gene was found and the correct start codon of the fungal transcription factor encoding gene was also established. The red boxes indicate the position of the exons.

It is interesting to notice that most of the genes analysed had a very high homology, of about 70% to 90%, to *Talaromyces marneffeii*. Even though both fungi belong to the same genus, *T. marneffeii* is a human pathogen, that causes respiratory, skin and systemic mycoses, and immunocompromised patients are particularly affected, therefore, this fungus has a very different

lifestyle compared to *T. versatilis* (Woo *et al.*, 2011; Yuen *et al.*, 2003). Another fungus that showed high sequence homology to *T. versatilis* is *Talaromyces stipitatus* (JGI Genome Portal), the closest fungus to *T. marneffeii*. Contrariwise, it is a non-pathogenic, but a saprophytic fungus found in soil and decaying plants. An example of this high homology between these three species is shown in Figure 4.2, with a multiple alignment of the XlnR proteins. For this protein, the *T. versatilis* protein shares 92% and 89% of identity with the *T. marneffeii* and *T. stipitatus* homologs, respectively. We also observed that there were regions that shared a considerable synteny between the three *Talaromyces* species. In Figure 4.3 the region flanking the Ku70-encoding gene is shown as an example. The *ku70* gene is involved in the non-homologous end joining (NHEJ) system that repairs double-strand breaks in DNA, and it is involved in the non-homologous recombination (see chapter 7). All the genes in this region showed high homology between the three organisms, but they also have a similar distribution.

Chapter IV

	1	70
XlnR T versatilis	(1)	MAQPSQT FGLDTLAE S SHYALEQLRLAREVDMN NSNS SGND FIKKDNKSAADPAGSRM QIMRSP LSLDAR
XlnR T marneffeii	(1)	MAQPSQT FGLDTLAE S SHYALEQLRLAREVDMN TDNNNE----FKKDNKAAADP VNSRMOIMRSP LSLDTR
XlnR T stipitatus	(1)	MAQPSQT SGLDTLAE S SHYALEQLRLAREL DMN INNS S NDS FIKKDSQS--DSAGSRM QIMRSP LSLDAR
Consensus	(1)	MAQPSQTPGLDTLAE S SHYALEQLRLAREVDMN NNNS FIKKDNKSAADPAGSRM QIMRSP LSLDAR
	71	140
XlnR T versatilis	(71)	AGIRKHSADT-AAVRRRISRACDQCQLRTRKCDGQSPCAHCTESGLSCEYARERKKRKGASKKDLAEAAA
XlnR T marneffeii	(67)	AGIRKHS TDA-ASVRRRISRACDQCQLRTRKCDGQSPCAHCTESGLSCEYARERKKRKGASKKDLAEAAA
XlnR T stipitatus	(69)	SIRKHSADTAAVRRRISRACDQCQLRTRKCDGQSPCAHCTESGLSCEYARERKKRKGASKKDLAEAAA
Consensus	(71)	AGIRKHSADT AAVRRRISRACDQCQLRTRKCDGQSPCAHCTESGLSCEYARERKKRKGASKKDLAEAAA
	141	210
XlnR T versatilis	(140)	KAAGGS-REPCTPGYETIPDQSSQLSVMASESDARLSQSRRSFSASQVADQOPGLASLRELAQQPPQTR
XlnR T marneffeii	(136)	KASGG-----TPVYEAVPDQSSQLSVMASENDARLSQSRRSFSASQVAGQOPGLASLRELAQQPPQTR
XlnR T stipitatus	(139)	KAAGGSFDSCTPGYETIPDQSSQMSVMTSESDARLDQSRRSFSASQVPOQ--SLSLRELAQQPPQTR
Consensus	(141)	KAAGG RD GTPGYETIPDQSSQLSVMASESDARLSQSRRSFSASQVA QOPGLASLRELAQQPPQTR
	211	280
XlnR T versatilis	(209)	SRQFF-FNMNSMAMNCGYGVQNVNDRSFIQMPDLRDLQPRSPSAIVPAGLNGFHDAYNMVHAP TINIYQY
XlnR T marneffeii	(200)	PRQFSAANMNSMTMNEGDVQNVNDRSFIQMPDLRDLQPRSPFAMVFPAGLNGFHDAYNMVHAP TINIYQY
XlnR T stipitatus	(207)	RPRQYFSPNMTPLMNCYGVQNVNDRSFIQMPDLRDLQPRSPSAIVPAGLNGFHDAYNMVHAP PSINIYQY
Consensus	(211)	RQFF PANMNSM MNGYGVQNVNDRSFIQMPDLRDLQPRSPSAIVPAGLNGFHDAYNMVHAPDVAVDINIYQY
	281	350
XlnR T versatilis	(278)	QYIQSGEDNANHFHTGLTPPVQSPGWLPLPAPSVGFQSLNMSGFTSTLKYVPLEP LPHIVSIIPQSLA
XlnR T marneffeii	(270)	QYPCSGEDASANHFHTGLTPPVQSPGWLPLTPS-GFPSLNMNGFSTLKYVPLEP LPHIASIIPQPLA
XlnR T stipitatus	(277)	PSQP-EDASANHFHTGLTPPVQSPGWLPLPAPSVGFPEINMSGFTSTLKYVPLEP LPHIVSIIPQSLA
Consensus	(281)	QYSQSGEDASANHFHTGLTPPVQSPGWLPLPAPSVGFPSLNMNGFTSTLKYVPLEP LPHIVSIIPQSLA
	351	420
XlnR T versatilis	(348)	CDLLDLYFASTSSSHVFPQSPYVVGHVFRKKSILHQTQPRTCSPALLASMLWVAQAQTS DAPFLTSPPSAR
XlnR T marneffeii	(339)	CDLLDLYFASTSASHVFPQSPYVVGHVFRKKSILHQTQPRTCSPALLASMLWVAQAQTS DAPFLTSPPSAR
XlnR T stipitatus	(346)	CDLLELYFASTSSSHVFPQSPYVVGHVFRKKSILHPTQPRTCSPALLASMLWVAQAQTS DAPFLTSPPSAR
Consensus	(351)	CDLLDLYFASTSSSHVFPQSPYVVGHVFRKKSILHQTQPRTCSPALLASMLWVAQAQTS DAPFLTSPPSAR
	421	490
XlnR T versatilis	(418)	GRVCQKLELTVNLLRPLIHGPTPGEASPNYAANAVINGVALGGFGVMSDQLGAQSSATGAVDDVATYIH
XlnR T marneffeii	(409)	GRVCQKLELTVGLLRPLIHGPTPGEASPNYAANAVINGVALGGFGVMSDQLSAQNSATGVDDVATYIH
XlnR T stipitatus	(416)	GRVCQKLELTVGLLRPLIHGPTPGEASPNYAANAVINGVALGGFGVMSDQLGAQSSATGAVDDVATYIH
Consensus	(421)	GRVCQKLELTVGLLRPLIHGPTPGEASPNYAANAVINGVALGGFGVMSDQLGAQSSATGAVDDVATYIH
	491	560
XlnR T versatilis	(488)	LATVVSASEYKASIRWMAAAWSLARELKLGRELPPTPSQPQSHDRDGNIDMEPKPSRDTSHVTEEBERE
XlnR T marneffeii	(479)	LATVVSASEYKASIRWMAAAWSLARELKLGRELPPTPSQPQSHDRDGNIDMEPKTSRDTSHITEEBERE
XlnR T stipitatus	(486)	LATVVSASEYKASIRWMAAAWSLARELKLGRELPPTPSQPQSHDRDGNVMEPKPSR-YQSVTEEBERE
Consensus	(491)	LATVVSASEYKASIRWMAAAWSLARELKLGRELPPTPSQPQSHDRDGNIDMEPKPSRDTSHVTEEBERE
	561	630
XlnR T versatilis	(557)	RRRIWLLYVMDRHLALCYNRPLTLLDKECEGLLQPMNDDVWHTGDFANAGYRRAGPSFECTGHGMFGYF
XlnR T marneffeii	(549)	RRRIWLLYVMDRHLALCYNRPLTLLDKECEGLLQPMNDDVWHTGDFANAGYRRAGPSFECTGHGMFGYF
XlnR T stipitatus	(555)	RRRIWLLYVMDRHLALCYNRPLTLLDKECEGLLQPMNDDVWHTGDFANAGYRRAGPSFECTGHGMFGYF
Consensus	(561)	RRRIWLLYVMDRHLALCYNRPLTLLDKECEGLLQPMNDDVWHTGDFANAGYRRAGPSFECTGHGMFGYF
	631	700
XlnR T versatilis	(627)	LPLMTILGEIVDLNHARNHPRFGIHFRTSGEWDSSHTEITRQLDVYEQSLREFETRHASLGIGNEGA--
XlnR T marneffeii	(619)	LPLMTILGEIVDLNHARNHPRFGIHFRTSGEWDSSHTEITRQLDVYEQSLREFETRHASLGIGSDGASG
XlnR T stipitatus	(625)	LPLMTILGEIVDLNHARNHPRFGIHFRTSGEWDSSHTEITRQLDVYEQSLREFEARTHASLGIGNEGA-A
Consensus	(631)	LPLMTILGEIVDLNHARNHPRFGIHFRTSGEWDSSHTEITRQLDVYEQSLREFETRHASLGIGNEGA A
	701	770
XlnR T versatilis	(695)	ADAGFNA-AAPTGDHVS PARSSTVGSRVNESLMQTKMVVAYGTYLMHVLHILLVGKWDPI SLLDDND
XlnR T marneffeii	(689)	ADAGFNA PVAAPTGDHVS PARSSTVGSRVNESLMQTKMVVAYGTYLMHVLHILLVGKWDPI SLLDDND
XlnR T stipitatus	(694)	PDTGFNA-AAPTGDHVS PARSSTVGSRVNESLMQTKMVVAYGTYLMHVLHILLVGKWDPI SLLDDND
Consensus	(701)	ADAGFNA AAPTGDHVS PARSSTVGSRVNESLMQTKMVVAYGTYLMHVLHILLVGKWDPI SLLDDND
	771	840
XlnR T versatilis	(764)	LWISSEAFITAMGHAVKAAEAASDILEYDPDLSFMPFFFGIYLLQGSFLLLLTADKLGQDADPSVVRACE
XlnR T marneffeii	(759)	LWISSEAFITAMGHAVKAAEAASDILEYDPDLSFMPFFFGIYLLQGSFLLLLTADKLGQDADPSVVRACE
XlnR T stipitatus	(763)	LWISSEAFITAMGHAVKAAEAASDILEYDPDLSFMPFFFGIYLLQGSFLLLLTADKLGQDADPSVVRACE
Consensus	(771)	LWISSEAFITAMGHAVKAAEAASDILEYDPDLSFMPFFFGIYLLQGSFLLLLTADKLGQDADPSVVRACE
	841	905
XlnR T versatilis	(834)	TIVRAHEACVVTLNTEYQRNFRKVMRSALAQVRGVPDDDFGEQQRRREV LALYRWTGDGSGLAL
XlnR T marneffeii	(829)	TIVRAHEACVVTLNTEYQRNFRKVMRSALAQVRGVPDDDFGEQQRRREV LALYRWTGDGSGLAL
XlnR T stipitatus	(833)	TIVRAHEACVVTLNTEYQRNFRKVMRSALAQVRGVPDDDFGEQQRRREV LALYRWTGDGSGLAL
Consensus	(841)	TIVRAHEACVVTLNTEYQRNFRKVMRSALAQVRGVPDDDFGEQQRRREV LALYRWTGDGSGLAL

Figure 4.2: Multiple sequence alignment of the XlnR proteins of *T. versatilis*, *T. marneffeii* and *T. stipitatus*. The identical amino-acids are indicated by the yellow background, the similar ones have a green background and the conserved amino-acids have a light blue background. The *T. versatilis* protein has a 92% and 89% identity with the *T. marneffeii* and the *T. stipitatus* XlnR proteins, respectively.

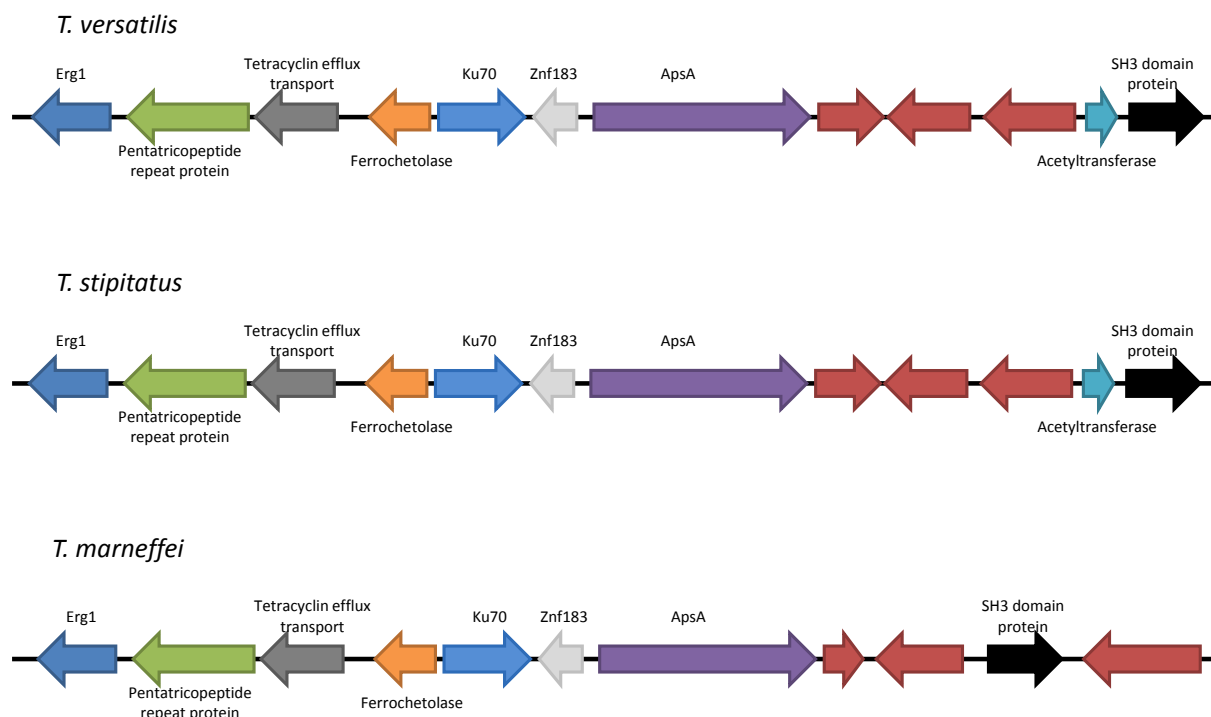


Figure 4.3: Comparison of a genomic region in *Talaromyces* genome sequences. Arrows that are the same colour represent genes that have closely related homologues across the three species. Red arrows represent genes coding for hypothetical proteins with unknown function. The protein name or functions are indicated above or below the arrow.

4.1.1. Glycoside hydrolases encoding genes

One of the focuses of this project was the analysis of the genes that code for glycoside hydrolases. Besides the tools mentioned before, the Pfam domains that are associated to each one of the GH families were also used to localize the genes of interest, and this proved to be a very helpful method. In total, we identified 170 genes encoding for GH enzymes, belonging to 32 different GH families relevant for the degradation of the plant polymers. In the CAZy classification there are 133 GH families that group a great variety of enzymes with hydrolytic capacity. It is therefore very likely that the number of genes that could be identified in *T. versatilis* genome is actually higher.

To better assess the genetic potential of *T. versatilis* for the degradation of the plant polymers, a comparison of the number of genes between *T. versatilis*, and four different *Aspergilli* (*A. niger*, *A.*

nidulans, *A. fumigatus* and *A. oryzae*), *Penicillium chrysogenum* and *Trichoderma reesei* (Andersen *et al.*, 2011; van den Berg *et al.*, 2008; Galagan *et al.*, 2005; Machida *et al.*, 2005; Martinez *et al.*, 2008; Nierman *et al.*, 2005; Pel *et al.*, 2007; Wortman *et al.*, 2009) was done. The total number of genes coding for the 32 GH families in each fungus shows that *T. versatilis* is one of the fungi with the highest number of genes, only overtaken by *A. oryzae* (Figure 4.4). *T. reesei* and *P. chrysogenum* are the fungi with the lowest number of GH-encoding genes. Surprisingly, *T. reesei* is one of the most important fungi for the industrial production of hydrolytic enzymes. The limited number of enzymes encoded by *T. reesei* seems to be enough for the degradation of the plant polymers (Martinez *et al.*, 2008). However, *T. versatilis* shows an important genetic potential for the degradation of complex polysaccharides.

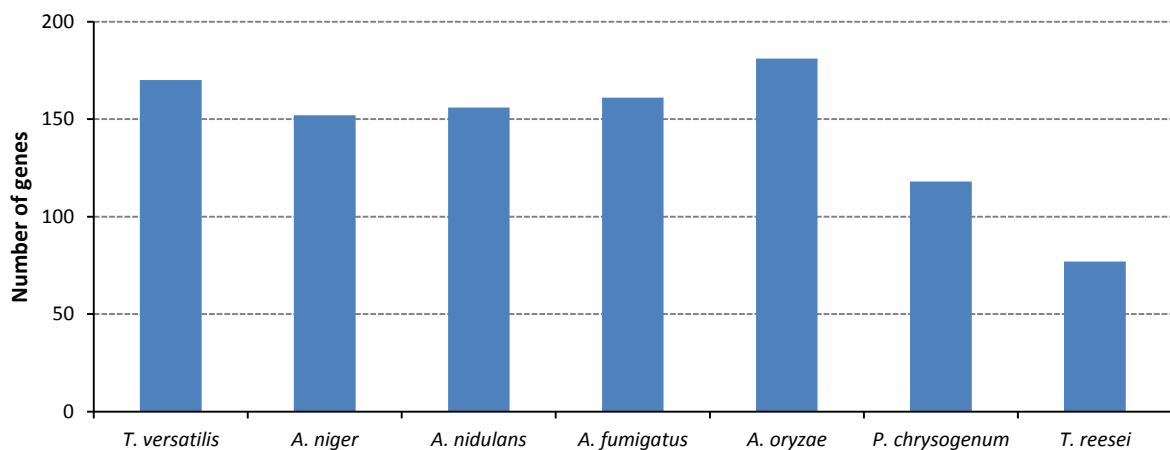


Figure 4.4: Total number of GH-encoding genes from 32 GH families in different filamentous fungi.

The Figure 4.5 shows the total number of genes from each of the 32 families for each fungus. Note that one family can be composed of different kinds of enzymes, the number of genes coding for each kind of enzymes in each fungus is detailed in the Appendix I. It is remarkable the overrepresentation of some of the families in *T. versatilis* compared to the other fungi. For instance, the arabinofuranosidases from the family GH54 and from the family GH62 have higher copy number in *T. versatilis* than in the other filamentous fungi studied. The endoxylanases of the family GH11 and

the α -glucuronidases from the family GH67 present a similar situation. The β -mannosidase encoding gene from the family GH2 is present in 6 copies in *T. versatilis*, only comparable to the copy number in *T. reesei* (see Appendix I). Also, *T. versatilis* has 20 copies of the β -glucosidase-encoding gene of the family GH3, whereas the other fungi have between 8 and 16 copies, being *T. reesei* the one with the lowest copy number (see the Appendix I for more details). On the other hand, some families are not represented at all, such as the families GH93 and GH95. The GH95 family is present in all the other fungi analysed, whereas the exoarabinanase of the family GH93 is also absent in *T. reesei* and *A. niger*.

It is interesting to notice that most of the families that have a higher gene copy number in *T. versatilis* encode enzymes that are responsible for the degradation of the hemicellulose. The cellulases encoding genes have a copy number equivalent to the other fungi. This may suggest that *T. versatilis* has the potential to degrade more complex substrates than just cellulose and particularly different kinds of hemicellulose.

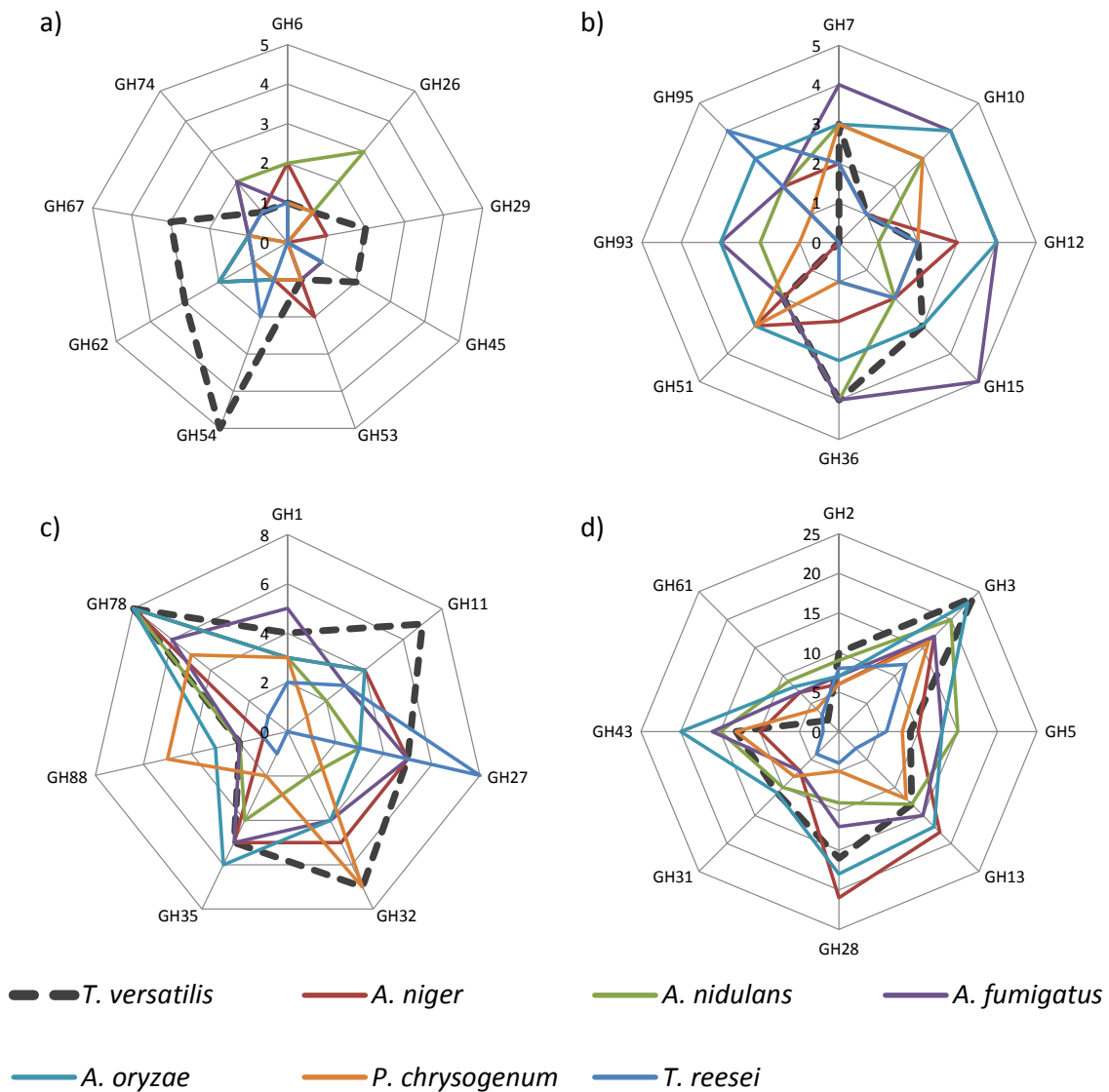


Figure 4.5: Number of genes encoding GH of different CAZy families in the 6 fungi analysed. The GH families have been separated in 4 different groups according to the average number of genes in each family: a) low number of genes, b) and c) middle number and d) high number of genes.

The degradation of plant cell wall polymers is not only carried out by the GH, but there are also other auxiliary proteins, that have no enzymatic activity and that may intervene in order to enhance the process by making the substrates more accessible for the actual hydrolases. The *T. versatilis* genome encodes eight different hydrophobic binding proteins, three of which are homologous to the

hsbA gene (Ohtaki *et al.*, 2006) and one to the *rolA* gene (Takahashi *et al.*, 2005), both described in *A. oryzae*. Also two genes encoding for putative swollenins were found (Table 4.1).

Table 4.1: List of genes coding for hydrophobins and swollenins in *T. versatilis*.

ORF in <i>T. versatilis</i>	Protein	Protein length	Pfam domain	Manual correction
Pf120c034g00170	Hydrophobic surface binding protein A (HsbA)	186 aa	PF12296	No
Pf120c037g00350	Hydrophobic surface binding protein A (HsbA)	173 aa	PF12296	No
Pf300c046g01130	Hydrophobic surface binding protein A (HsbA)	290 aa	PF12296	No
Pf330c014g00230	Hydrophobin	171 aa	PF12296	No
Pf120c037g00340	Hydrophobin	318 aa	PF12296	No
Pf120c082g02440	Hydrophobin	356 aa	PF01185	Yes
Pf120c084g00760	Hydrophobin (RolA)	152 aa	PF01185	No
Pf300c020g00160	Hydrophobin	176 aa	PF01185	Yes
Pf170c005g00140	Swollenin	503 aa		No
Pf300c046g01280	Swollenin	453 aa		Yes

4.1.2. Transcriptional regulators involved in the GH production and secretion

Homologues of transcriptional regulators that are well known in other fungi (see section 1.3 for details on transcriptional regulation of GH genes in fungi) are also found in the genome of *T. versatilis* (Table 4.2). Homologues to *creA*, *creB* and *creC*, which are involved in the carbon catabolite repression, have been identified. The activators of cellulases and hemicellulases that respond to the presence of xylose, arabinose, cellulose and mannobiose: *xlnR*, *araR*, *clrA* and *manR* respectively, are also present. We could find the eight genes involved in the Pal pathway, which is regulated by the pH of the environment, namely, *palA*, *palB*, *palC*, *palF*, *palH*, *palI*, *pacC* and *vps32/snf7*. The *areA* gene that controls the expression of the cellulases in response to the availability of a nitrogen source, was identified in the genome. And finally, we located the *hacA* regulator, responsible for the

repression of the expression of cellulases and hemicellulases genes during the Unfolded Protein Response (UPR).

Table 4.2: List of known transcriptional regulators found in the *T. versatilis* genome and their function.

Stimuli	Response	Transcriptional regulator	ORF in <i>T. versatilis</i>	Manual correction
Glucose	Repression of genes involved in the use of alternative carbon sources	CreA	Pf300c054g00700	No
		CreB	Pf120c033g00560	No
		CreC	Pf160c019g00010	Yes
Cellulose	Repression of xylanase and cellulase genes	ACEI	Pf330c016g00640	No
Xylose	Induction of xylanase and cellulase genes	XInR	Pf160c026g00520	No
Arabinose	Induction of arabinolytic genes	AraR	Pf120c012g00470	No
Cellulose	Induction of cellulases genes	ClrA	Pf170c010g00810	Yes
β -1,4 mannobiose	Induction of cellulases and mannanases genes	ManR	Pf120c037g00840	No
pH	Induction of alkaline expressed genes	PalA	Pf170c016g00980	Yes
		PalBory	Pf200c039g00470	Yes
		PalC	Pf200c012g01510	No
		PalF	Pf200c029g01060	No
		Pall	Pf200c060g00620	No
		PacC	Pf200c012g01510	Yes
		Vps32/Snf7	Pf170c013g01730	Yes
Nitrogen starvation	Induction of cellulase genes	AreA	Pf390c013g00160	Yes
Unfolded proteins	Unfolded protein response	HacA	Pf270c010g01350	Yes

The ACEI transcription factor has been described as a repressor of the cellulase and xylanase genes in *Trichoderma reesei* as the $\Delta aceI$ mutant produces cellulases at a higher level than the wild type strain (Aro *et al.*, 2003). Interestingly, a homologue of the *aceI* repressor is also present in the *T. versatilis* genome. The ORF encoding the ACEI protein in *T. versatilis* lacks a fragment of approximately 750 bp. The protein produced is not complete and it might not be functional, as already shown for the truncated form of the *aceI* gene from *T. reesei*, which encodes a protein that lacks the N terminal region and cannot act as a repressor of the cellulase and xylanase gene. The ACEI proteins in other fungi are between 750 and 850 amino-acids, whereas the encoded protein in *T. versatilis* is only 449 amino-acids and the fragment missing is on the N terminal (Figure 4.6). The annotation of *aceI* was manually curated, but no sequence coding for the missing part was found in the vicinity of the gene. The lack of the N terminal region of the *aceI* homolog from *T. versatilis*, and therefore the lack of one of the main cellulase repressors, may explain why this strain is such a good enzymes producer. To probe this theory, a comparison of the DNA region coding for the *aceI* in the wild type strain with the evolved industrial strain was done. The region coding for this gene was amplified in the wild type *T. versatilis* strain, and the amplicon was sequenced. No difference was found between the sequences of the *aceI* locus in the industrial strain and the wild type strain, meaning that this gene is also truncated in the wild type strain. The *acell* regulator could not be found in the genome of *T. versatilis*; this is not surprising since this regulator is characteristic to *Trichoderma* and was not found in other fungi, including the Aspergilli.

4.1.3. Secretion pathway

Although the knowledge about the secretory pathway in fungi is still limited, it is considered to be similar to the one in yeast and higher eukaryotes. The main difference would be that in fungi, the secretion occurs mainly in the apical area of the hyphae (Girard *et al.*, 2013). *T. versatilis* is not the exception to the rule. As it is shown in the Table 4.3 no major differences were found in the number or nature of the genes involved in the secretion pathway when comparing our fungus of interest to *A. nidulans*.

The genes involved in the different parts of the secretion process were identified by BLAST using the *A. nidulans* genes as models and are listed in the Table 4.3. The genes coding for the signal recognition, signal peptidases and translocation into the ER were identified, all of which are involved in the transport of the proteins into the ER. However, homologous to the *SEC65* and *SRP14* genes could not be found in *T. versatilis*. All the genes coding for the proteins from the chaperone cycle and protein folding process were found, except for the homologous of the *SIL1* gene. And finally, the UPR and ER associated degradation (ERAD) genes were also recognized. The percentage of identity at the amino-acid level shown in the table corresponds to the comparison of the *T. versatilis* proteins with those of *A. nidulans*. The level of identity between the proteins from both organisms is considerably high, always above 40% and reaching 90% in some cases, suggesting that this pathway is highly conserved.

The failure to find the *SRP14* and *SIL1* genes amongst the analysed group does not necessarily mean that they are absent in the *T. versatilis* genome. The fact that the genome is still at a contig level, means that there is an important number of gaps, and it is possible that the genes in question are encoded elsewhere than in the known parts of the genome. If that is not the case, it would be

possible to consider that the functions accomplished by those proteins are replaced by other non characterised proteins in *T. versatilis*.

Table 4.3: List of genes involved in the secretion pathway in *T. versatilis*, compared to *A. nidulans*.

<i>S. cerevisiae</i> Gene Code	<i>A. nidulans</i> ORF	<i>T. versatilis</i> ORF	Amino-acid Identity (%)	Manual correction
Entry into the ER				
Signal Recognition				
<i>SEC65</i>	AN10106.3	Pf230c011g00180	69	No
<i>SEC65</i>	AN0643.3	NF		
<i>SRP14</i>	AN4580.3	NF		
<i>SRP21</i>	AN0068.3	Pf100c002g01370	49	No
<i>SRP54</i>	AN8246.3	Pf120c084g00810	89	No
<i>SRP68</i>	AN4043.3	Pf200c028g00430	55	No
<i>SRP72</i>	AN2014	Pf360c030g00990	57	No
Signal Peptidases				
<i>SEC11</i>	AN10354.3	Pf170c011g00970	79	No
<i>SPC1</i>	AN11133.3	Pf200c059g00370	74	No
<i>SPC2</i>	AN1525.3	Pf300c045g00610	59	No
<i>SPC3</i>	AN9460.3	Pf120c082g02320	61	No
Translocation into ER				
<i>SRP101</i>	AN6627.3	Pf170c008g00650	77	No
<i>SRP102</i>	AN5819.3	Pf160c009g00450	55	No
<i>SBH1</i>				
<i>SBH2</i>	AN0417.3	Pf200c024g00520	90	Yes
<i>SSS1</i>	AN4589.3	Pf290c003g00380	78	No
<i>SEC61</i>				
<i>SSH1</i>	AN7721.3	Pf120c085g00270	96	No
<i>SEC62</i>	AN6269.3	Pf390c011g01790	70	No
<i>SEC63</i>	AN0834.3	Pf120c019g00380	71	No
<i>SEC66</i>	AN1442.3	Pf240c022g00760	66	No
<i>SEC72</i>	AN10987.3	Pf160c027g00550	64	No
Processes in the ER				
Chaperone cycle				
<i>KAR2</i>	AN2062.3	Pf160c018g00120	86	No
<i>LHS1</i>	AN0847.3	Pf160c010g00500	63	No
<i>SIL1</i>	AN7272.3	NF		
Protein folding: SS bonds				
<i>PDI1</i>	AN0075.3	Pf100c002g01620	61	No
<i>EPS1</i>	AN5970.3	Pf300c056g00240	64	No
<i>EUG1</i>	AN7436.3	Pf160c014g00190	72	No
<i>MPD1</i>	AN0248.3	Pf200c039g00410	60	No
Associated with folding				
<i>CNE1</i>	AN3592.3	Pf120c033g00760	79	No
<i>CPR5</i>	AN4467.3	Pf390c013g01120	80	Yes
<i>ERJ5</i>	AN5770.3	Pf390c013g00370	59	Yes
<i>ERO1</i>	AN1510.3	Pf120c013g00640	63	No
<i>ERV2</i>	AN3759.3	Pf120c029g00120	55	Yes
<i>FMO1</i>	AN8206.3	Pf200c020g00410	43	No
<i>SCJ1</i>	AN6170.3	Pf160c008g00070	72	No
Quality Control				

Unfolded Protein Response				
CDC48	AN7254.3	Pf160c025g01490	92	No
GCN4	AN3675.3	Pf120c036g01030	42	No
HAC1	AN9397.3	Pf270c010g01350	67	Yes
IRE1	AN0235.3	Pf360c038g00070	66	Yes
KAR2	AN2062.3	Pf160c018g00120	86	No
ORM1				
ORM2	AN1933.3	Pf120c082g01550	91	No
PTC2	AN1358.3	Pf170c013g01080	68	Yes
SCJ1	AN6170.3	Pf160c008g00070	72	No
TRL1/RLG1	AN1296.3	Pf170c013g00610	61	No
ER-associated Degradation				
ASI1				
ASI3	AN2442.3	Pf160c025g02850	57	Yes
DER1	AN12122	Pf330c014g00550	74	No
	AN3307.3	Pf300c055g00790	64	No
Doa10	AN7704.3	Pf120c079g00510	69	No
DOA4	AN2072.3	Pf160c017g00130	68	Yes
DOA10/SSM4	AN6136.3	Pf100c002g00300	48	No
DSK2	AN5451.3	Pf350c004g01770	61	No
HRD1	AN1488.3	Pf240c019g00070	58	No
HRD3	AN0810.3	Pf120c019g00500	63	Yes
NPL4	AN0295.3	Pf200c030g00970	71	No
RSP5	AN1339.3	Pf170c013g00980	85	Yes
UBC1				
RAD6/UBC2	AN2212.3	Pf170c005g00670	77	No
UBC4				
UBC5	AN2761.3	Pf230c019g00490	95	No
UBC6	AN10874	Pf160c028g00080	41	No
UBC7	AN8258.3	Pf120c084g00310	90	No
UBR1	AN3923.3	Pf100c010g00460	64	No
UFD1	AN0329.3	Pf370c002g00660	79	No
UFD2	AN10556.3	Pf170c016g01790	68	No
YOS9	AN1461.3	Pf240c022g00070	49	No

4.2. Improvement of the genomic annotation using the RNA-seq data

During this thesis, RNA-seq data was produced as part of the analysis of the transcriptome of *T. versatilis* (See chapter 6). The sequencing was performed using the SOLiD technology, and 43 to 52 million reads were successfully aligned to the genome of *T. versatilis* for each of the conditions tested. As described in the Materials and Methods section (chapter 2), the reads obtained from the RNA sequencing were first filtered. The data was then used to reconstruct the transcripts that are expressed in the conditions studied. Thereafter, the transcripts could be aligned to the genomic sequence using TopHat. This is a program that aligns RNA-Seq reads to a genome and allows the identification of exon-exon splice junctions. In parallel, different softwares were used to predict the

position of the ORFs: Cufflinks, to assemble the aligned RNA-Seq reads into transcripts and for a *de novo* prediction of genes, GeneMark was used, this is an algorithm that determines the protein-coding potentials and the probability of protein coding in a DNA fragment in the six possible reading frames.

The new predictions also gave about 18,000 genes. Interestingly, only 9,000 had RNA-seq reads aligned. From the other 9,000 genes, some are probably real genes not expressed during the transcriptomic experiment or not detected during the sequencing, but we could reasonably expect that many of them are not real genes. Besides the quantitative debate, a considerable improvement to the gene annotation was done eliminating the mistakes that were discussed in the previous sections. This procedure was carried out by the bioinformaticians from the Deep-seq facility at the University of Nottingham. The new annotation is stored in a “.gtf” (General Transfer Format) file that is accompanied by the FASTA file containing the genomic sequence. The advantage of the .gtf file is that it can be easily modified if a mistake is detected and a manual curation is carried out. In this way, the manual correction can be immediately incorporated into the *T. versatilis* annotation.

4.3. Discussion

Having a correct annotation in a genome is essential for the downstream functional genomics works. The errors made by gene prediction algorithms during the automatic annotation, like missing genes, incorrect gene boundaries, misplaced or missing exons and wrongly merged genes can jeopardize any attempts to accurately predict the metabolic potential of the organism and the development of any experimental approach (Cerqueira *et al.*, 2014). The automatic annotation of the *T. versatilis* genome presented all kinds of problems, as highlighted by the manual curation of this annotation that was initiated at the very beginning of this thesis work.

The manual curation of the *T. versatilis* genome was an important task, considering that there are over 18,000 genes annotated and many of them are probably wrong. When we compare the

genome of *T. versatilis* with the Aspergilli, for instance, we realize that the size of the genome is similar, but the number of genes is much bigger in *T. versatilis*. While focusing our manual curation on the three groups of genes of interest, the similarity of *T. versatilis* sequences with Aspergilli and Penicillia was undeniable. But the highest homology was found with *T. marneffeii* and *T. stipitatus*. The first one is a human pathogen (Woo *et al.*, 2011; Yuen *et al.*, 2003), and the second is a saprophytic fungus. Many of the glycoside hydrolases encoding genes were found to have high homology to *T. marneffeii* which was somehow surprising. Zhao *et al.* (2013) showed that *T. marneffeii* has an important number of hydrolytic enzymes, and therefore a good plant cell wall degrading capacity. This fungus is not an obligated parasite and, when it is not invading a host, it can live as a saprophytic fungus, which could explain the need for so many GH-encoding genes. *T. stipitatus* also had an important level of homology with *T. versatilis*; this was expected since *T. stipitatus* is a saprophytic fungus that survives without any host. The high sequence homology was not the only thing that was in common amongst the three fungi; regions of synteny were also detected, where the position of the genes was identical and the genes shared high homology, as mentioned before. This means that the syntenic region underwent none or very low rearrangement events and it also indicates how closely related these three species are. *T. versatilis* used to belong to the *Penicillium* genus, and now is part of the *Talaromyces* genus. The re-classification was done based on the partial ribosomal RNA gene clusters (part of 18S small subunit RNA gene, internal transcribed spacer 1 (ITS1), 5.8S ribosomal RNA gene, internal transcribed spacer 2 (ITS2), part of 28S large subunit ribosomal RNA gene) and *btuB* gene sequences (Bridge & Buddie, 2013). The synteny regions and high sequence homology observed during the manual curation between the *T. versatilis*, *T. marneffeii* and *T. stipitatus* and presented here support the re-classification of *T. versatilis*.

Even though only 32 GH families were analysed during the manual annotation, *T. versatilis* showed a very high genetic potential for the degradation of the plant polymers. *T. versatilis* presented a wide variety of enzymes with the capacity to degrade a large range of substrates, which may be linked to the ability that this fungus showed to grow over complex substrates (see chapter 3).

The large number of hydrolase-encoding genes in *T. versatilis*, that was higher than most of the other fungi analysed, included genes from various families. The need for a large variety of enzymes is explained by the fact that a synergistic action of all those hydrolases is necessary for an efficient degradation of the complex polymers that are present in the plant cell wall. *T. reesei* is one of the most important industrial fungi, still is one of the fungi with the least GH-encoding genes. In comparison to *T. versatilis*, the number of GH of *T. reesei* seems poor. It is possible that this fungus has the right combination of enzymes that act in synergy for the full degradation of the lignocellulosic material.

The hemicelluloses arabinoxylan, arabinan and arabinogalactan are decomposed by the enzymes that belong to the GH51 and GH54 families mainly (Zhao et al., 2013). A similar number of genes coding for GH51 family enzymes was observed in all the fungi studied. *T. versatilis* has a significantly higher amount of genes coding for GH54 and GH62 enzymes, in comparison to the other fungi. Some of the arabinofuranosidase of this family were already studied in our team from a biochemical point of view (Guais, 2009; Guais et al., 2010; De La Mare, 2014). The reason why *T. versatilis* needs so many arabinofuranosidases is not yet clear. This fungus also has a bigger number of xylanases from family GH11. Moreover, some of the GH families, in particular the families involved in the degradation of the hemicelluloses, are overrepresented in *T. versatilis*, although the expression of all these genes might not be at the same time and some of them could be necessary only in particular circumstances, like stress conditions. *T. versatilis* is clearly well prepared for the degradation of hemicelluloses. The regulation of the expression of the GH-encoding genes will be studied in the chapters 6 and 7.

Genes encoding for hydrophobin and swollenins were also found in the *T. versatilis* genome. Although there is still no clear evidence of their involvement in the plant cell wall degradation, it is possible that they have an indirect role in the process, by enhancing the action of other enzymes and making the substrate more available for the decomposition. Swollenins have been shown to be

involved in the interaction with crystalline cellulose and enhance its degradation (Chen *et al.*, 2010; Wang *et al.*, 2011). Similarly, hydrophobins were described to be involved in the interaction with hydrophobic substrates and enhance their degradation by other enzymes (Ohtaki *et al.*, 2006; Takahashi *et al.*, 2005). The number of hydrophobin genes is surprisingly large in *T. versatilis*, but they might be involved in different processes in the cell besides the interaction with hydrophobic substrates, like reducing the surface tension or covering the conidia cell wall. The presence of these auxiliary proteins encoding genes and the GH genes indicates that this fungus has all the necessary tools for a proficient degradation of complex and diverse substrates with different physicochemical characteristics.

The transcriptional regulators involved in the control of the expression of the GH-encoding genes in response to different stimuli were identified in the *T. versatilis* genome. For most of them, no major differences were found in the sequences, and the homology with the known transcription factors from other fungi was very important. The only exception was ACEI, a repressor that was described in *T. reesei* (Aro *et al.*, 2003; Saloheimo *et al.*, 2000). Its homologous gene in *T. versatilis* encodes a protein that lacks the N terminal region and, even though the annotation of this gene was manually curated, the missing part could not be found. The comparison of the evolved industrial strain with the original, wild-type strain showed no difference whatsoever in the nucleotide sequence of the locus, indicating that the partial deletion is not a product of the mutagenesis process used to obtain the industrial strain. However, having a dysfunctional cellulases and hemicellulases repressor may be one of the reasons why *T. versatilis* is a very efficient enzymes producing fungus.

It has been shown that the hyper-cellulolytic fungus *T. reesei* RUT C30 lacks a 85 kb fragment from its genome compared to the wild type strain (Ilmén *et al.*, 1996; Seidl *et al.*, 2008), which leads to a truncated *cre1* gene and therefore to the inactivation of the main cellulases repressor. Likewise, there must be a difference between the *T. versatilis* wild-type strain and the industrial strain that makes the second one a much better enzymes producer than the first one, but based on sequence

analysis of the *creA* locus, it seems that the *T. versatilis* industrial strain is endowed with a completely functional carbon catabolite repression system. Targetted deletion of the locus will nevertheless allow answering this question. Besides, a comparative analysis of the genomes of these two strains (industrial *versus* original, wild type strain) could be considered, if the genome of the wild type strain is sequenced.

Another aspect studied was the secretion pathway. During the secretion process, which is fundamental for the discharge of the hydrolytic enzymes into the media, the ER coordinates the folding and some post-translational modifications of proteins that reside in, or pass through, the endomembrane system. The large flux of proteins being translocated into the ER generates a need to enhance the efficiency of protein folding and transport as well as the quality control of the synthesized proteins (Guillemette *et al.*, 2007). A great number of genes are involved in the entire process, which also includes the UPR and the ERAD. The sequence homology between *A. nidulans* and *T. versatilis* secretion genes was outstandingly high, even though they belong to different genera and are not very close in the phylogeny. Moreover, all of the genes from *A. nidulans* included in this study were found in *T. versatilis*, with only 3 exceptions. This leads us to think that the secretion pathway has been very well conserved in spite of the speciation events occurred during the evolution and separation of these two fungal species.

The annotation problems mentioned before became more and more evident and the manual curation of the genome turned out to be a gigantic time-consuming task. Moreover, when we carried out the whole-genome transcriptomic study (see chapter 6), the annotation of the genome was a big problem for the proper alignment of the reads and quantification of the transcript level for each gene. At this point, we considered to use these RNA-seq data to improve the genomic annotation. The RNA-seq data can allow the identification of all expressed transcripts, the complete and contiguous mRNA sequence from the transcription start site to the transcription end can be identified and the multiple alternatively spliced isoforms distinguished (Grabherr *et al.*, 2011). The

large amount of data provided by the transcriptomic study gave us the possibility to improve the structural annotation of the genome, still in an automated way. Other algorithms for the prediction of genes, like GeneMark, were also used. The combination of all these tools resulted in an improved annotation that will simplify the future functional genomics studies.

5. Study of the transcriptome of *Talaromyces versatilis*

5.1. Genome wide transcriptome study using RNA-seq of *Talaromyces versatilis* exposed to lignocellulose

5.1.1. First RNA-seq results analysis

The conditions used in this experiment are three physiologically different conditions: the first one is a repressing condition, where the Carbon Catabolite Repression (CCR) is active since glucose is present; the mycelia is then exposed to a lignocellulosic material which induces a set of genes for the degradation of the substrate; finally, one could expect to set up a repressing condition again with the incubation of the mycelium in presence of the lignocellulosic material and glucose at the same time. Practically, *T. versatilis* conidia were first incubated for 48h on TvMM containing 1% (w/v) of glucose as sole carbon source (hereafter called: Glucose 48h). The mycelium was then filtered and transferred to fresh TvMM with 1% wheat straw as carbon source, and it was incubated for 24h (Straw 24h). The wheat straw used in this experiment is composed of $37 \pm 1.7\%$ of cellulose, $32 \pm 1.2\%$ of hemicellulose and $22 \pm 0.1\%$ lignin (Delmas *et al.*, 2012) and it was ball milled and autoclaved before use. Finally, glucose was added to the culture at a final concentration of 1%, and the flasks were incubated for 5h more (Straw+Glucose 5h). Mycelia samples were taken at each step for RNA extraction, semi-quantitative PCR and RNA sequencing as described in the Materials and Methods section (see chapter 2).

In Figure 5.1 a semi-quantitative PCR is shown over 3 genes: the *abf-B2* (Pf200c012g01410), the *cbhI* (Pf360c030g01900) and *cbhII* (Pf200c030g01540). This experiment allowed the validation of the RNA samples obtained with the experimental design presented before, prior to the actual sequencing. The tubulin encoding gene (Pf120c035g00270) was used as a control, even though now we know that it is not the best choice for a reference gene (see the next section), for the purpose of

this experiment it was good enough. No expression for any of the GH-encoding genes was detected on Glucose 48h. But a strong activation during the incubation on wheat straw was detected, as a matter of fact, the cDNA of the cellobiohydrolases is noticeable after 15 cycles and the arabinofuranosidases, after 20 cycles. In the Straw + Glucose 5h condition there is still cDNA from the three GH genes, but in a lower concentration, since the amplicons are only visible after 20-25 cycles.

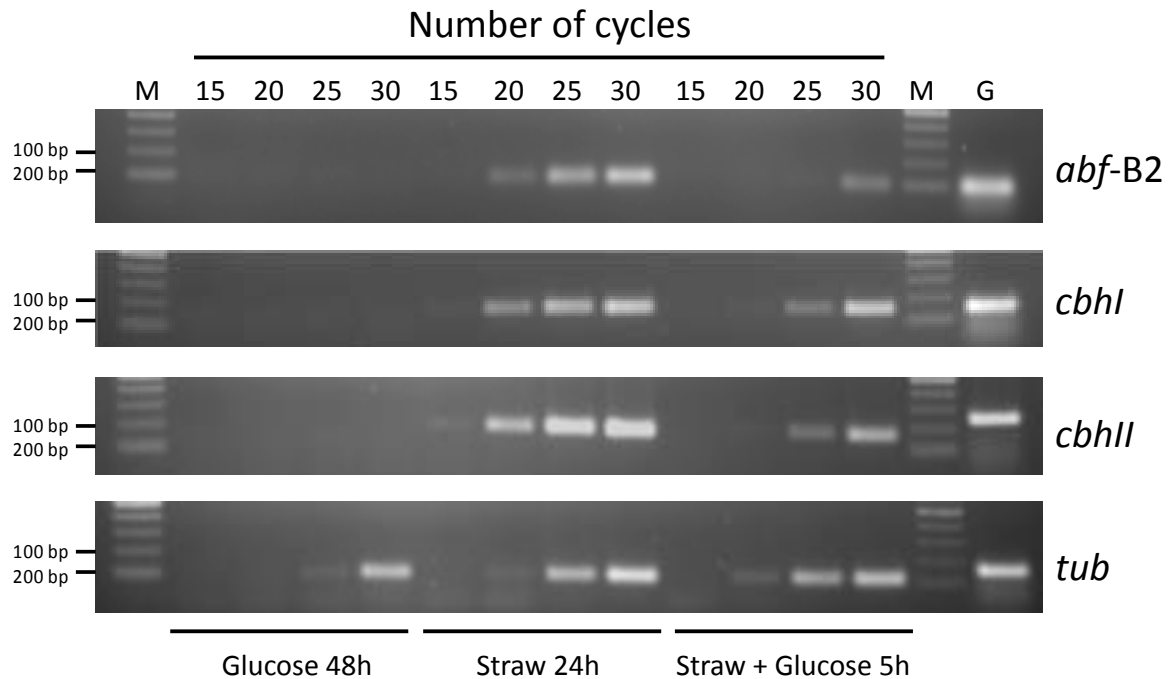


Figure 5.1: Semi quantitative PCR. Electrophoresis in agarose gel (1.5%) of the PCR products obtained using the primers designed for the *abf-B2*, *cbhI*, *cbhII* and *tub* encoding genes. For the 3 conditions - Glucose 48h, Straw 24h and Straw + glucose 5h - and each pair of primers, samples were taken after 15, 20, 25 and 30 cycles of PCR. The hydrolases genes are only expressed in the second and third condition, with higher intensity in straw in the absence of glucose. The tubulin gene is expressed in every condition. M: 100 bp DNA ladder. G: *T. versatilis* genomic DNA control.

After the sequencing and the alignment of the reads to the genome, the RPKM value was calculated for each gene in each of the biological replicates. The reproducibility of the experiment was extremely high; the R squared value was always above 0.95 within the three biological replicates.

A correlation of all the expression values (RPKM) obtained for all the genes in the genome between the Glucose 48h and Straw 24h condition is shown in Figure 5.2 a. The scatter plot shows a central group of genes that are well correlated between the two conditions. These genes fall over the linear regression line (black central line) and they are not differentially expressed between the two conditions; they have fairly the same RPKM value for both conditions. But the genes that are detached from the central group and outside the prediction interval (red lines) correspond to the genes whose expression level changes from one condition to the other. The significantly differently expressed genes, determined by the DEGseq, and that will be discussed below, are represented by the green circles. The genes on top of the regression line have very low expression values on Glucose 48h and show a high induction on Straw 24h. Similarly, there are a few points below the regression line that were expressed on Glucose 48h and repressed on Straw 24h. In the same way, the correlation of the expression values of all the genes between Straw 24h and Straw+Glucose 5 h is represented in the Figure 5.2 b. An important group of genes are spread below the regression line, they correspond to the genes strongly up-regulated in Straw 24h that are down-regulated with the addition of glucose.

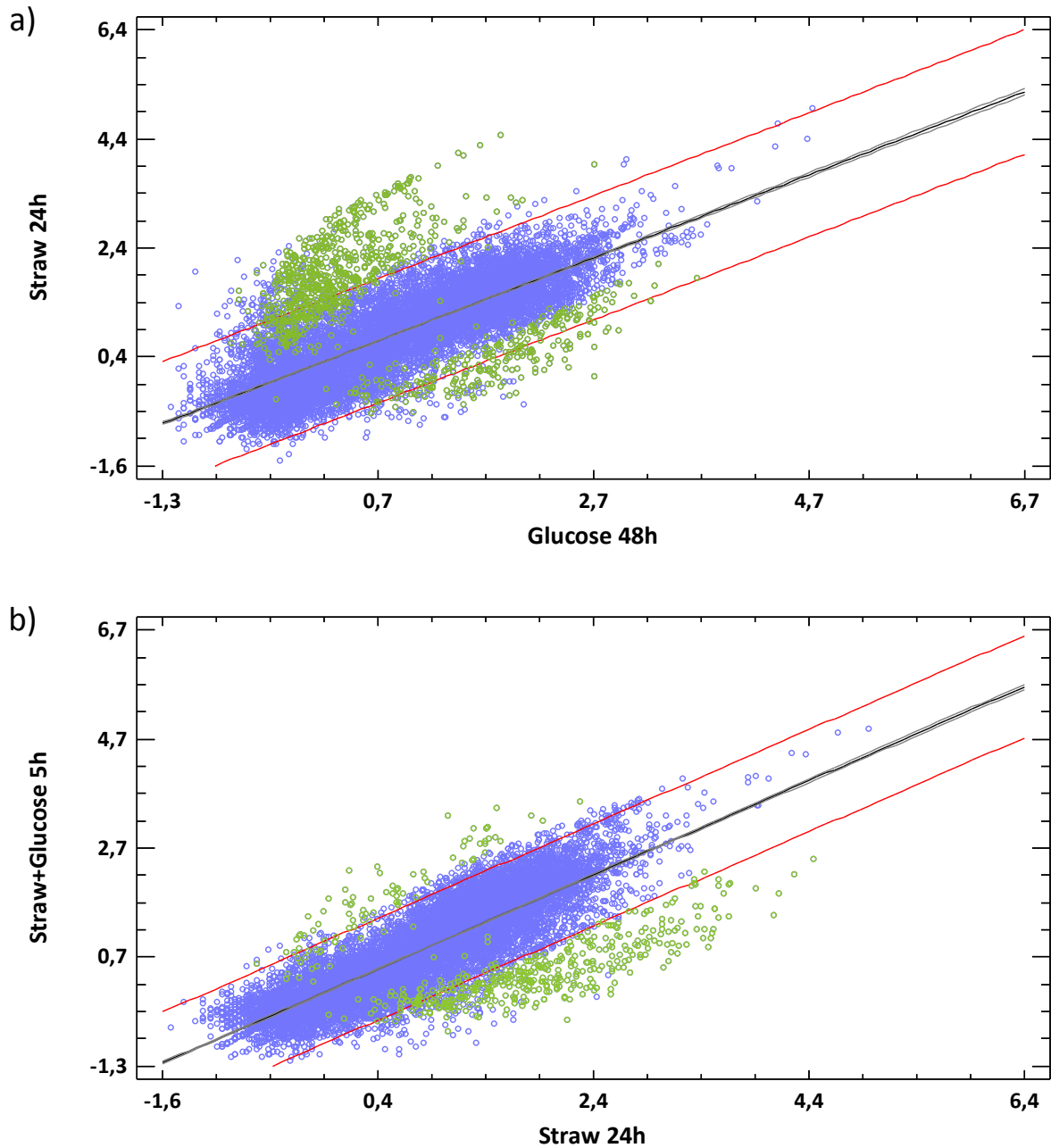


Figure 5.2: Correlation of the expression of all genes in *T. versatilis* between the different conditions tested. Each circle represents a gene. The RPKM values were log₁₀-transformed. The black line represents linear regression. Grey lines and red lines are the 95% confidence intervals and 95% prediction intervals, respectively. The green circles correspond to the significantly differently expressed genes. a) Expression values in Glucose 48h vs. Straw 24h. b) Expression values in Straw 24h vs. Straw + glucose 5h.

The significantly differently expressed genes were determined using the DEGseq pipeline (Wang *et al.*, 2010b) (see chapter 2). A total of 926 genes were significantly differentially expressed between

the conditions Glucose 48h and Straw 24h. From which 657 were up-regulated when the mycelia was exposed to wheat straw, and the other 269 were down-regulated. When glucose was added to the straw culture, 499 genes were significantly differentially expressed, 407 being repressed and only 92 being induced. Surprisingly, only 314 genes were found to be differentially expressed, up or down regulated, in both conditions, as it is shown on Figure 5.3. More precisely, within the 657 genes that were induced in Straw 24h, only 265 turned off when glucose has been added to the culture. The other way around, from the 269 genes repressed under Straw 24h, not more than 49 were up-regulated again when glucose is supplemented to the media.

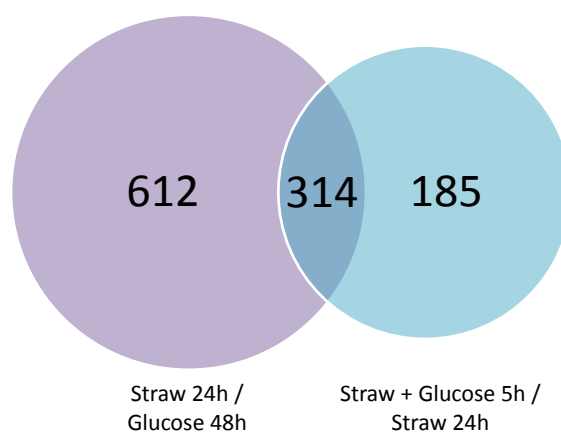


Figure 5.3: Venn diagram showing the number of differentially expressed genes in the Straw 24h in relation to the Glucose 48h condition (violet), and Straw + Glucose 5h in relation to the Straw 5h condition (blue).

The group of gene induced in Straw 24h includes GH encoding genes, accessory proteins and transcription factors. These genes will be discussed in the following sections. We also found 51 genes encoding putative transporters and 12 genes encoding for putative esterases. The list of these genes and their expression levels can be found in the Appendix IV.

5.1.1.1. *Expression of the glycoside hydrolases genes*

The *T. versatilis* genome encodes at least 172 glycoside hydrolases from 32 different CAZy families. From those genes, 81 were differentially expressed, which represents the 48% of the total GH-encoding genes. A list of the GH genes, with their annotation, RPKM values for the 3 conditions and differential expression can be found in the Appendix V. As it is shown in Figure 5.4, most of the GH families analysed bear differentially expressed gene, with the exception of families GH29, GH32 and GH51 that bear no gene presenting differential expression. In fact, the genes that belong to these families have a rather low level of expression, which could be considered null, in every condition. The huge GH13 gene family only exhibited 1 differentially expressed (DE) gene amongst the 13 identified members. Besides these exceptions, all the other families presented an important fraction of genes, if not all, which were strongly regulated when the fungus is exposed to wheat straw.

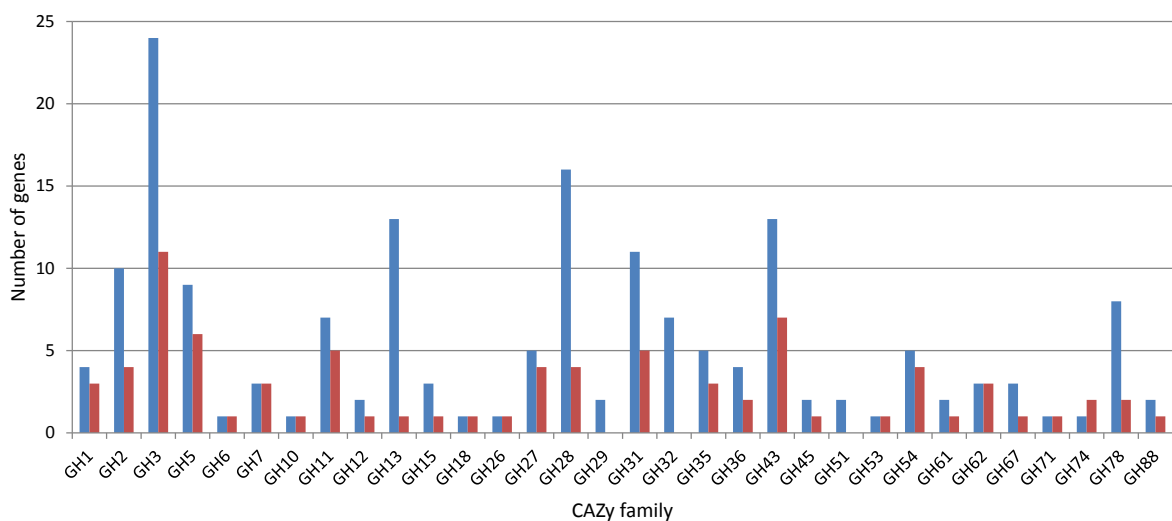


Figure 5.4: Number of GH-encoding genes for each CAZy family analysed (blue bars) and number of genes of each family differentially expressed (dark red bars).

In general terms, most of the GH-encoding genes, including those that were differentially expressed, followed the same pattern: the expression was strongly enhanced in Straw 24h and then reduced once more with the addition of the glucose, but the inhibition in the third condition was

never as strong as the inhibition observed in Glucose 48h. The fold-change values of the GH-encoding genes between the Straw 24h and Glucose 48h and between Straw + Glucose 5h and Straw 24h were calculated. The distribution of the \log_2 of the fold-change values is shown in Figure 5.5.

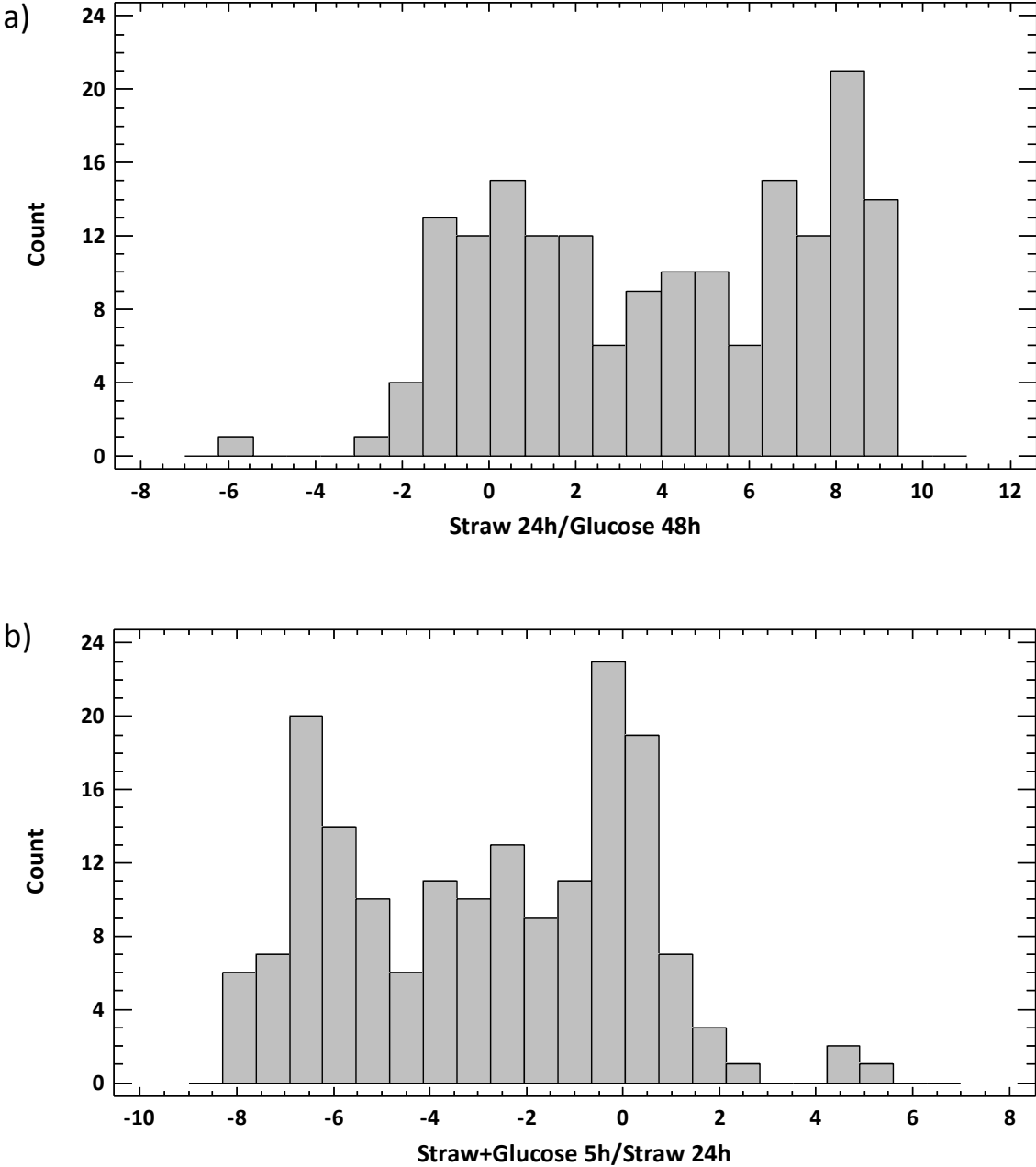


Figure 5.5: Histogram of frequency versus the fold-change values (\log_2). a) Distribution of the ratio Straw 24h/Glucose 5h. Most of the genes are activated in Straw 24h. b) Distribution of the ratio Straw + Glucose 5h/Straw 24h. Most of the genes are repressed in the third condition.

To better illustrate the massive activation of GH encoding genes, we analysed the total amount of RPKM associated to these gene families, relative to the total amount of RPKM that was obtained from the RNA-seq run. As can be seen in Figure 5.6, the expression of the genes coding for GH enzymes represents only 0.15% of the total transcriptome of *T. versatilis* after 48h of incubation on glucose. This is consistent with the carbon catabolite repression state typically observed in the presence of easily metabolized sugars, such as glucose (Aro *et al.*, 2005; de Vries & Visser, 2001). After being transferred to medium with wheat straw, *T. versatilis* strongly activates a set of GH-encoding genes in order to cope with the lack of an easily metabolisable carbon source and the degradation of the complex lignocellulosic material. In this condition, the mRNA that corresponds to the GH genes represents 13.22% of the entire transcriptome. This massive shift of the transcriptome is remarkable, since the 13.22% of the total mRNA in the cell comes from only 172 genes, which represent 0.91% of the total number of genes in the genome. When glucose is added to the culture, the amount of mRNA corresponding to the GH genes only represents 0.28% of the transcriptome, which confirms that an important repression of the expression of this group of genes has occurred, even though the polysaccharides were still present in the medium.

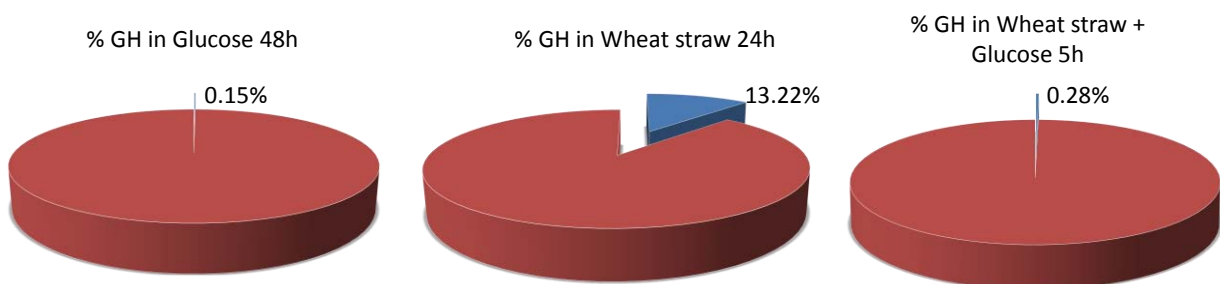


Figure 5.6: Percentage of transcripts corresponding to GH-encoding genes (blue) in relation to the total mRNA in the cell, set to 100%, for each of the three conditions studied.

Similar situations were observed in *A. niger* and *T. reesei* (Delmas *et al.*, 2012; Ries *et al.*, 2013), in which 13.76% and 19.13% of the total mRNA correspond to CAZy-encoding genes, respectively. It

is important to note that the carbohydrate esterase (CE) and polysaccharide lyase (PL) encoding genes are also included in these percentages. The inclusion of those genes was not possible in *T. versatilis* due to the low quality of the genome annotation, which would have required a manual selection of the genes. However, the amount of mRNA from GH-encoding genes in *T. versatilis*, probably under-estimated, is comparable to these two fungi.

In Figure 5.7 the percentages of mRNA of each GH family have been represented for the three conditions tested, bearing in mind that these mRNA represent a small part of the total mRNAs in the cell (Figure 5.6). In presence of glucose, the enzymes from the families GH3, GH13 and GH31 are the most abundant, being the GH31 the one with the highest amount of mRNA. These families encode for β -galactosidases, α -amylases, and α -glucosidases. Extremely low levels of mRNA for enzymes necessary for the deconstruction of plant polymers are present, many of which have such low RPKM value that can be considered as repressed genes.

Upon exposure to wheat straw, xylanases from GH11 family are highly up-regulated, as they represent 38% of the mRNA corresponding to all the GH-encoding genes. The cellobiohydrolases from family GH7 and, to a lesser extent, the families GH3, GH5, GH6 and GH10 also represent a considerable percentage of the mRNA in the cells. All together these families account for over 75% of the total mRNA of the GH-encoding genes. This is a remarkable finding, specially taking into account that within those families only 32 genes show a significant up-regulation in presence of wheat straw. These enzymes, that include xylanases, β -glucosidases, β -xylosidases, glucanases, arabinofuranosidases and cellobiohydrolases, are most likely the necessary activities for the saccharification of the wheat straw. The increased diversity in the enzyme panel in this condition is a reflection of the complexity of the substrate, the fungus needs different kind of enzymes to degrade the cellulose, but also the side chains and the backbone of the hemicelluloses present. Both *A. niger* and *T. reesei* showed an important increase in the diversity of the enzymes expressed in this condition. In these fungi, GH11 was also amongst the highly expressed genes, along with GH7

(Delmas *et al.*, 2012; Ries *et al.*, 2013). The strong induction of the GH11 xylanases is in accordance with the studies carried out in *T. reesei* and *N. crassa* that demonstrated that the expression of these enzymes is induced in presence of xylose or arabinose (Herold *et al.*, 2013; Sun *et al.*, 2012), both sugar are indeed present in the wheat straw and are probably released during the 24h of polymers deconstruction. The former GH61 family was also amongst the highly expressed genes in the two fungi, but it only counts for 1% of the GH mRNA in *T. versatilis*. Another interesting feature to take into account is the fact that, even though the GH54 family presents a much higher number of genes in *T. versatilis* than in the other fungi (see chapter 4), the proportion of mRNA corresponding to this family is not outstandingly high, just 3%, and most of the mRNA correspond to one of the genes of this family, the *abf-B2*, the other four genes are also induced, but to a lesser extent. This result is consistent with the observations made by Guais *et al.* (2010) of *T. versatilis* exposed to an industrial mimicking medium strongly expresses *abf-B2*, followed by *abf-B3*. In the case of wheat straw, *abf-B1* is the second most expressed gene and *abf-B3* does not show such a strong induction.

When glucose is supplemented to the mycelia growing on wheat straw, another important change in the transcriptome is observed. As mentioned before, the percentage of mRNA corresponding to the GH-encoding genes is importantly reduced, but also, the set of GH genes expressed in this condition changes one more time. As expected from the partial return to the repressed state, the proportions of RNA of each class of GH did not recover the pattern that was observed during growth on glucose. The GH11 and GH7 became less important in the last condition probably due to the carbon catabolite repression, and the the proportion of transcripts from families GH3, GH13 and GH31 increased.

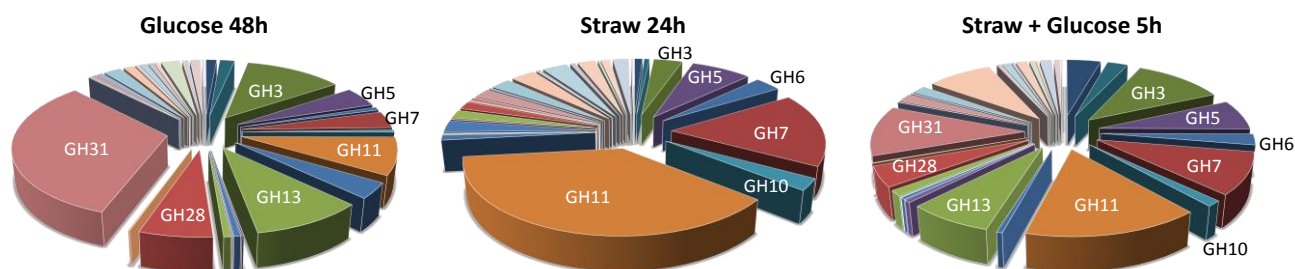


Figure 5.7: Proportion of mRNA of each GH family from the total GH mRNA for each condition of the transcriptome study.

To illustrate this, the expression patterns of genes from the families GH11, GH7, GH6, GH5 and GH3 were represented in Figure 5.8. A gene from each family is shown as an example. These results are consistent with the previous findings and confirm the trend of expression of the GH-encoding genes observed on the semi-quantitative PCR.

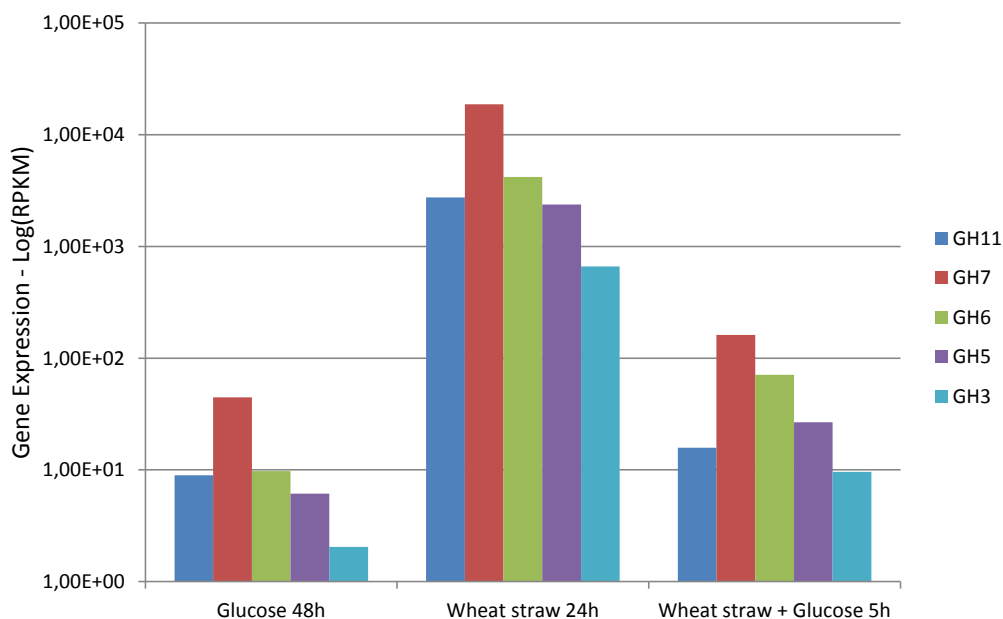


Figure 5.8: Expression pattern of genes encoding GHs from families 11, 7, 6, 5 and 3 in the three conditions: Glucose 48h, Straw 24h and Straw + Glucose 5h. The expression values are represented in a logarithmic scale.

5.1.1.2. Accessory proteins encoding genes induced by wheat straw

The genome of *T. versatilis* holds several genes that encode for hydrophobic binding proteins and also swollenins. The expression of five out of eight genes encoding for hydrophobic binding proteins was strongly induced in the Straw 24h condition. The expression levels of this group of genes is shown in Table 5.1. Four of the differentially expressed genes present the Pfam12296 domain, which is homologous to that of the HsbA from *A. oryzae* (Ohtaki *et al.*, 2006), but only three of these proteins have high homology to HsbA. The fifth differentially expressed gene presents the Pfam01185 hydrophobin domain. An induction of genes coding for hydrophobic binding proteins upon exposure to wheat straw was also reported in *A. niger* and *T. reesei* (Delmas *et al.*, 2012; Ries *et al.*, 2013). It was proposed for these fungi that these proteins can adsorb to the hydrophobic surfaces and promote its degradation. It was indeed suggested that the hydrophobins from different *Trichoderma* species enhance the degradation of poly(ethylene terephthalate) (PET), a synthetic polymer that presents a hydrophobic surface (Espino-Rammer *et al.*, 2013). Likewise, the hydrophobic binding proteins from *A. oryzae* have a role on the degradation of the synthetic polyester PBSA (Ohtaki *et al.*, 2006; Takahashi *et al.*, 2005). Specially, the *rolA* gene from was induced when *A. oryzae* is incubated on a medium with PBSA as sole carbon source, and the protein allowed the adhesion of the fungal hyphae to the PBSA surface (Takahashi *et al.*, 2005). The *rolA* homologous gene in *T. versatilis* showed a very low transcript level and was not differentially expressed. The function of this gene in *T. versatilis* and its expression pattern under a broader spectrum of conditions remains to be determined.

The two genes coding for swollenins found in the genome of *T. versatilis* were also strongly induced in the Straw 24h condition. Swollenins are proteins supposed to loosen the cellulose microfibrils and make them more available for the action of the hydrolytic enzymes (Saloheimo *et al.*, 2002). These proteins, along with the hydrophobic binding proteins, might be indirectly involved in the degradation of the plant polymers by facilitating the access to the substrate to other hydrolytic enzymes.

Table 5.1: Expression values (RPKM) of the hydrophobins and swollenin encoding genes. Differential expression (DE) is indicated in the last column.

Protein	Protein length	ORF in Pf	Pfam domain	RPKM Glucose 48h	RPKM Straw 24h	RPKM Straw + Glucose 5h	DE
HsbA	186 aa	Pf120c034g00170	PF12296	1,64	816,62	6,89	Yes
HsbA	173 aa	Pf120c037g00350	PF12296	31,43	12167,47	29,07	Yes
HsbA	290 aa	Pf300c046g01130	PF12296	11,07	2760,95	175,65	Yes
Hydrophobin	171 aa	Pf330c014g00230	PF12296	1,23	1,98	6,15	No
Hydrophobin	318 aa	Pf120c037g00340	PF12296	4,87	2326,73	15,75	Yes
Hydrophobin	356 aa	Pf120c082g02440	PF01185	1,99	3,1	1,38	No
Hydrophobin - RoIA	152 aa	Pf120c084g00760	PF01185	0	43,26	8,61	No
Hydrophobin	176 aa	Pf300c020g00160	PF01185	2,4	517,13	1471,51	Yes
Swollenin	503 aa	Pf170c005g00140		7,76	3393,86	38,2	Yes
Swollenin	453 aa	Pf300c046g01280		2,72	1192,71	45,24	Yes

5.1.1.3. Transcription factors

From the transcription factors inventoried during the genome annotation and that were of interest for the characterization of the regulatory networks controlling the GH encoding genes, only two genes were significantly differently expressed, namely *xlnR* and *clrA*. Both of them were up-regulated when the fungus was exposed to the lignocellulosic material. The detailed list of transcription factors, with the RPKM values for each condition and the differential expression are in Table 5.2. The fact that these 2 transcription factors are strongly up-regulated is not surprising, since they are necessary for the induction of the GH-encoding genes that will degrade the complex substrate to which the fungus is confronted in this condition and respond to the presence of xylose, in the case of *XlnR* and cellulose, in the case of *ClrA* (Coradetti *et al.*, 2012; van Peij *et al.*, 1998a).

The HacA regulator is slightly induced in Straw 24h, even though it is not considered to be significantly differently expressed by the DEGseq statistical tests. The HacA regulator is responsible for the activation of the Unfolded Protein Response (UPR), it will enhance the expression of the chaperones and foldases encoding genes in order to cope with the stress produced by the overloaded secretory pathway (Aro *et al.*, 2005; Pakula *et al.*, 2003). As it has been suggested for *A. niger*, a small increase in the *hacA* expression might be enough to provoke a differential expression in the UPR target genes, since even in conditions that do not represent a stress there is a small amount of HacA (Mulder *et al.*, 2006).

Table 5.2: Expression values (RPKM) of the transcription factors encoding genes. Differential expression (DE) is indicated in the last column.

Transcriptional regulator	ORF in Pf	RPKM Q10	Glu48	RPKM Straw24 Q10	RPKM StrawGlu5 Q10	DE
CreA	Pf300c054g00700	15,59		22,58	15,4	No
CreC	Pf160c019g00010	20,12		37,55	7,87	No
CreB	Pf120c033g00560	33,93		102,12	30,18	No
Ace1	Pf330c016g00640	292,77		188,6	134,74	No
XlnR	Pf160c026g00520	10,77		168,13	19,5	Yes
AraR	Pf120c012g00470	5,51		30,9	6,9	No
ClrA	Pf170c010g00810	1,52		49,12	6,86	Yes
ManR	Pf120c037g00840	3,41		5,46	1,85	No
PalH	Pf160c025g01260	20,44		10,55	18,38	No
PacC	Pf170c011g00070	16,07		35,06	13,51	No
PalA	Pf170c016g00980	29,59		30,25	28,1	No
PalF	Pf200c029g01060	38,19		8,23	21,13	No
PalBory	Pf200c039g00470	33,17		36,49	25,69	No
PalC	Pf200c012g01510	16,61		11,14	20,3	No

Pall	Pf200c060g00620	44,42	23,9	41,51	No
Vps 32/snf 7	Pf170c013g01730	78,9	78,64	60,69	No
AreA	Pf390c013g00160	25,59	18,34	37,99	No
HacA	Pf270c010g01350	181,82	688,5	177,27	No

Besides the two significantly differently expressed genes *xlnR* and *clrA*, there are 13 more genes encoding for transcription factors that are significantly differently expressed in the conditions studied. These genes are either transcription factors that are not directly related to the expression of the GH-encoding genes or transcription factors with unknown function. One of them codes for a homologous of the BrlA transcription factor, and another one for a homologous of the AlcR regulator. Amongst the remaining genes there are 9 genes that encode for transcription factors containing a Zn(II)₂Cys₆ binuclear cluster domain and 2 genes that encode for proteins that have a C₂H₂ zinc finger domains. The expression of all these genes is induced from 8 fold and up to 104 fold by the wheat straw substrate. The function of all these genes is still unknown, but further studies should be considered since their involvement in the degradation of the lignocellulose may be essential. The full list of significantly differently expressed transcription factors is given in Appendix VI.

BrlA is a transcription factor that contains a C₂H₂ zinc finger domain that is expressed upon conidiation signals and that is necessary for the conidiation process to start (Adams *et al.*, 1998). It has been shown that BrlA mediates the response to the nitrogen stress in *A. fumigatus*, since the deletion mutant was insensitive to nitrogen starvation but it responded normally to carbon starvation (Twumasi-Boateng *et al.*, 2009). The increased expression of this transcription factor in *T. versatilis* in the Straw 24h condition, where it reaches a fold change of 103.9, may suggest that the fungus is under a nutritional stress. The fact that the polymers present in the wheat straw are not easily metabolized and that the fungus needs to invest a great amount of energy to induce the GH-

encoding genes that will be capable of degrading the complex substrate may be the stress signal that enhances the expression of *brlA*, although nitrogen is not the lacking nutrient in the media.

The AlcR regulator controls the expression of genes associated to the alcohol metabolism – including the alcohol dehydrogenase and the aldehyde dehydrogenase encoding genes – and it is a protein that has a Zn(II)₂Cys₆ binuclear cluster domain. In *A. nidulans*, the expression of these genes is under the control of CreA, therefore when glucose is present the alcohol metabolism genes are repressed. But once the carbon catabolite repression is released, the genes from alcohol metabolism are expressed under the AlcR control, which also activates its own promoter (Mathieu & Felenbok, 1994; Mathieu *et al.*, 2005). In the case of *T. versatilis*, the transfer from a medium with glucose as carbon source to a medium with a complex lignocellulosic material induces the release of the carbon catabolite repression of the genes controlled by CreA. The de-repression of this set of genes allow expression of enzymes that begin the degradation of the substrate, the monomers released from the first degradation will act as signals for the full activation of the specific target genes, as it was suggested by Delmas *et al.* (2012) in *A. niger*. It is likely that the *alcR* gene, which may also be under the control of CreA is simply de-repressed in Straw 24h, although there is no clear need for the alcohol metabolism genes.

5.1.1.4. Antisense transcription

Natural antisense transcripts have already been described in several organisms, including fungi. They are non-coding RNA, transcribed from the opposite DNA strand to a coding strand and they are supposed to have a regulatory function (Faghihi & Wahlestedt, 2009). This kind of transcripts were described in *A. niger* (Delmas *et al.*, 2012; Novodvorska *et al.*, 2013), *Ustilago madis* (Donaldson & Saville, 2013), *T. reesei* (Ries *et al.*, 2013). The strand specific reads from the RNA-seq were mapped against the genome, and the antisense reads that fell into the annotated ORFs for each condition were used to calculate the antisense RPKM for each gene. The strand specific reads could be visualized using the IGV software. The sense and antisense reads are differentiated by distinctive

colours (Figure 6.8). In the example presented in Figure 5.9, the gene is encoded in the Crick strand (lower DNA strand) and it is composed by 7 exons. In the Glucose 48h condition, the reads are mapped to the Watson strand (upper DNA strand) and are therefore antisense (pink reads). Upon exposure to wheat straw, the transcription switches and the reads map to the Crick strand (blue reads). Some anti-ense reads are still detected in this condition. Also, in the first condition there is no splicing in the reads, but the splicing events clearly appear in the Straw 48h condition. The addition of glucose in the third phase of the experiment, switches again the sense of the transcription to antisense.

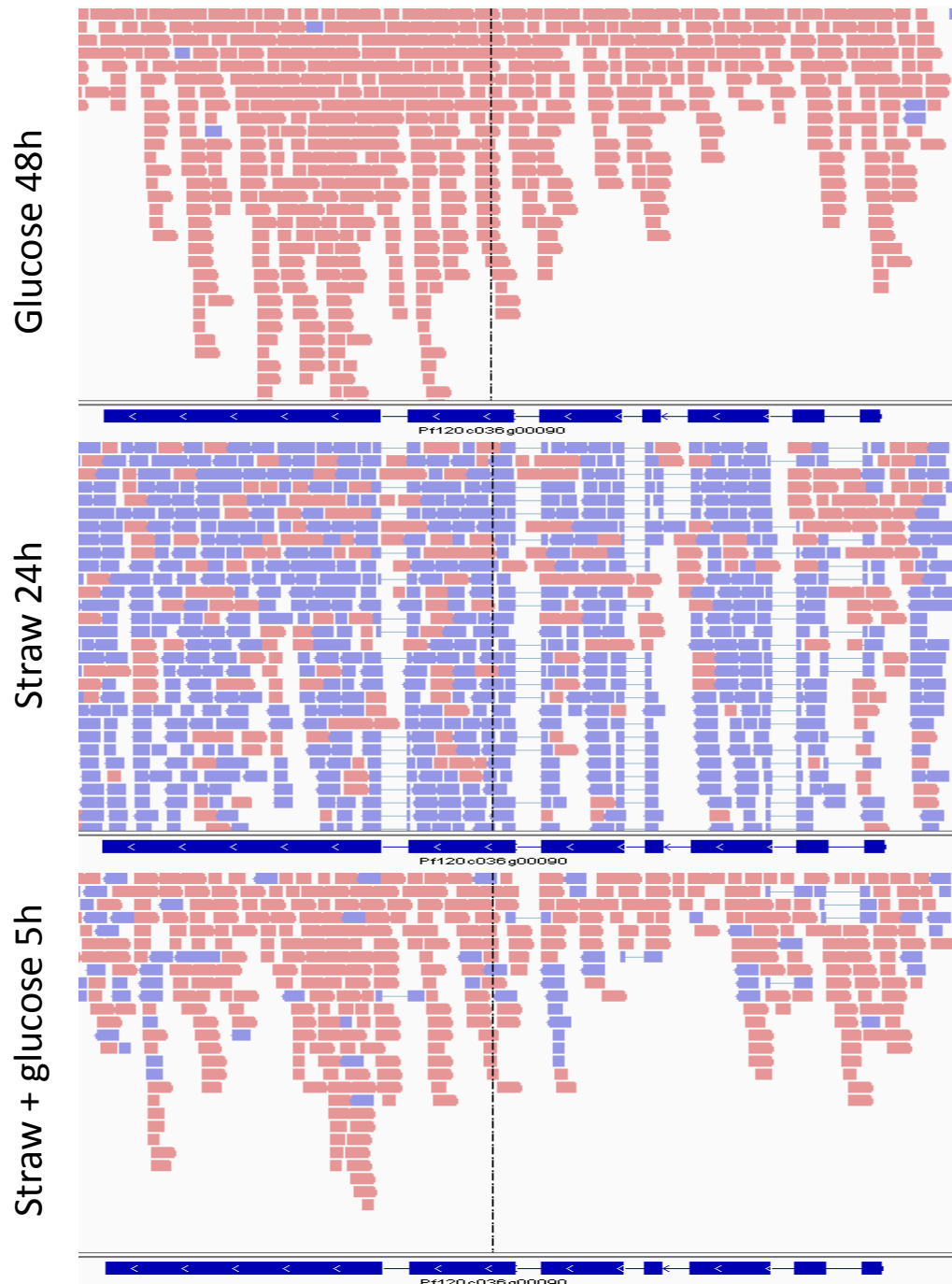


Figure 5.9: IGV visualization of the sense (blue) and antisense (pink) reads aligned to the genomic sequence. The diagram of the gene, which corresponds to a sugar transporter, is represented by the blue boxes. The full length gene is composed of 7 exons. The direction of the gene is indicated by the white arrows within the blue boxes. The introns splicing is indicated by the blue lines between two blocks of RNA transcripts.

A ratio between the sense (S) and the antisense (AS) expression values for each gene in each condition was calculated. When plotting the ratios obtained for Straw 24h versus the ratios for

Glucose 48h (Figure 5.10 a) we observe a large group of genes on the lower left quadrant that correspond to the genes that have a majority of sense transcripts in both conditions. In the lower right quadrant, there are genes that have mainly antisense transcription in Glucose 48h, but that switch to sense transcription in Straw 24h. Similarly, in the upper left quadrant there are the genes that have mainly sense transcription in Glucose 48h and antisense transcription in Straw 24h. Unexpectedly, there is an important group of genes that are plotted in the upper right quadrant; these are genes that have mainly antisense transcription in both conditions. It is possible that these genes are only transcribed in the antisense in these conditions and that they need other kind of signals to switch to the sense transcription. However, since the annotation of the genome used for this RNA-seq analysis has a rather low quality, the possibility of a wrong annotation has to be considered. In Figure 5.10 b the correlation between the ratios of Straw 24h versus Straw + glucose 5h is represented. The genes that switch from antisense transcription in Straw 24h to sense transcription in Straw + Glucose 5h are in the lower right quadrant. And those that switch from sense transcription in Straw to antisense in Straw + Glucose 5h are in the upper left quadrant. It is noticeable that the number of dots in this quadrant is much lower than in the lower right one.

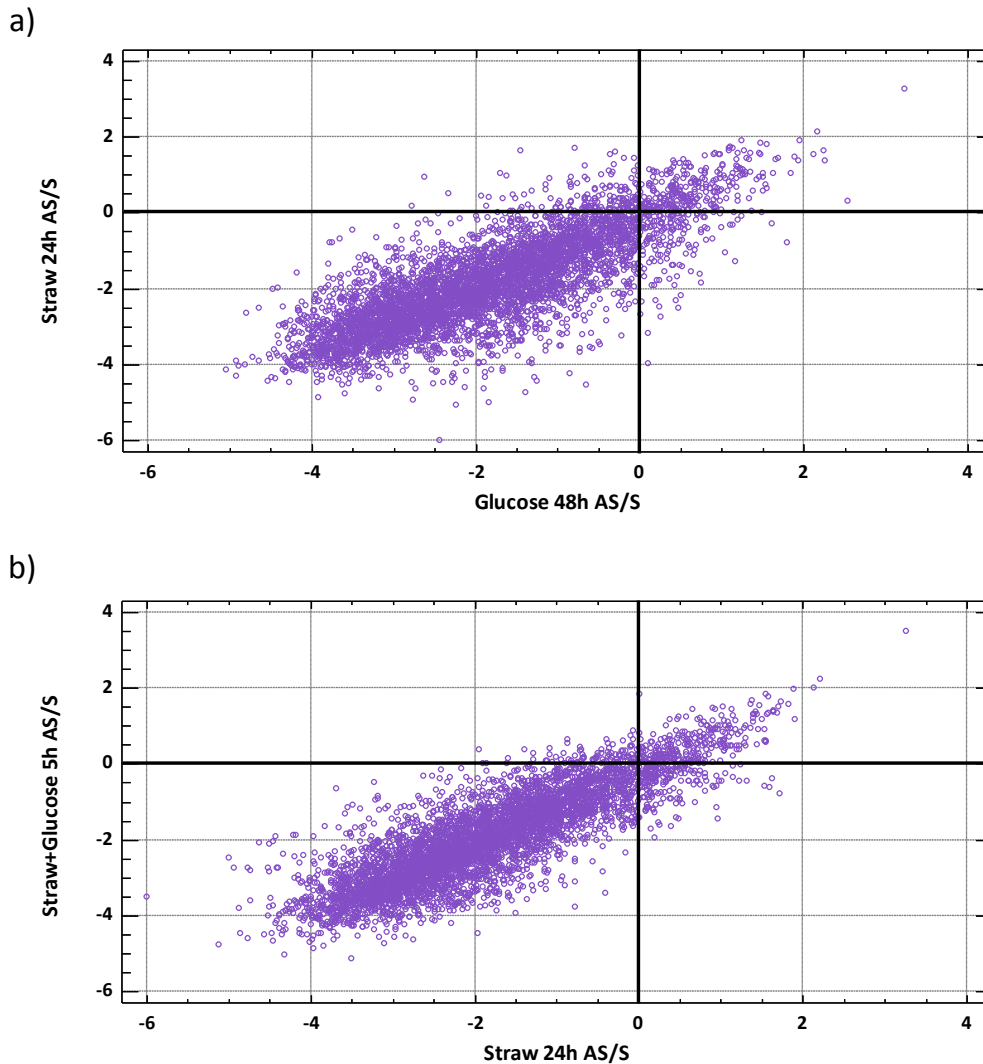


Figure 5.10: Scatter plot of the AS/S ratios for the three conditions. a) correlation of log of the AS/S ratios in Glucose 48h vs. Straw 24h. b) correlation of the log of the AS/S ratios of Straw 24h vs. Straw + Glucose 5h.

In total, 317 genes present above 5 RPKM of antisense transcription in at least one of the conditions studied in this work in *T. versatilis*. The RPKM cut-off value of 5 was set arbitrarily. None of the GH-encoding genes identified in *T. versatilis* so far have antisense transcription. Neither the hydrophobic binding proteins encoding genes, nor the transcription factors that are involved in the expression of the GH-encoding genes show antisense transcription. In fact, 96 out of 317 are genes annotated as “hypothetical protein” and most of them have a very low transcript level in most of the conditions, or have antisense expression in the 3 conditions. Other 89 genes are annotated as

“putative uncharacterized protein”. Amongst the remaining genes there are transcription factors with unknown function, transporters and dehydrogenases/reductases. The detailed list of genes with their sense and antisense RPKM values and annotation can be found in the Appendix VII. The function of the antisense transcripts is still unknown in filamentous fungi, but it has been suggested that they might be involved in the impediment of the expression of unnecessary proteins, exerting a post-transcriptional regulatory function (Novodvorska et al., 2013).

5.1.2. RNA-seq remapping to the new genome annotation

The new genomic annotation done by the Deep-seq facility at the University of Nottingham allowed improving the annotation of the genome of *T. versatilis* (see Chapter 4). From this point it was possible to re-map the RNA-seq data obtained previously and re-calculate the RPKM for each gene in each condition. Once again, the R squared obtained when comparing the biological replicates was always above 0.9, which indicates a very good reproducibility. Interestingly, from the 18,349 genes predicted in the new annotation, only 10,789 genes have at least one read mapped in at least one of the conditions tested. Some of the 7,560 genes left might be actual ORFs, but it is very likely that many of them are inaccuracies from the automatic annotation. Although the results have not changed dramatically in terms of number of genes, the quality of the mapping has been importantly improved.

This time the significantly differently expressed genes were determined using the DESeq pipeline (Anders & Huber, 2010). Between the Glucose 48h and the Straw 24h conditions there are 4,204 significantly differently expressed genes, from which 2,427 genes are induced in Straw 24 and the other 1,777 genes are repressed in this condition. When comparing the Straw 24h with the Straw + Glucose 5h condition, there are 2,754 significantly differently expressed genes, 1,163 of which are up-regulated in Straw + Glucose 5h and 1,591 are down-regulated in the same condition. There are 2,371 genes that show a differential expression in the two conditions. The number of significantly differently expressed genes is therefore considerably higher compared to the previous analysis

(4,204 versus 926 DE genes during the shift from glucose to wheat straw). However, the DESeq analysis is considered to be more accurate than the DEGseq, because the negative binomial distribution mathematical model used in the DESeq is more representative of the read counts data than the Poisson distribution used by the DEGseq.

The gene ontology (GO) is a classification that provides an ontology of defined terms representing gene product properties. There are three domains: cellular component, molecular function and biological process (<http://www.geneontology.org/>). The first one corresponds to the parts of a cell or its extracellular environment. The second category describes the elemental activity of a gene product at a molecular level. And the third category is the sets of molecular events pertinent to the functioning of an integrated living unit. The GO enrichment analysis consists in assessing which GO attributes are enriched within the group of genes being studied. A GO enrichment analysis was performed with the new data. The analysis was applied to the significantly differently expressed genes in Straw 24h vs. Glucose 48h and in Straw + Glucose 5h vs. Straw 24h, but the up-regulated and down-regulated genes were treated separately. The results are listed in Table 5.3.

Within these results it is clear that upon exposure to wheat straw there is an important enhancement of the GO terms related with the hydrolysis, cellulase, xylanase activities, polysaccharides catabolism and also sugar transporters. These GO terms decrease when glucose is added to the medium. The fungus does an important investment to produce the necessary enzymes for the full degradation of the lignocellulosic substrate in this condition, but glucose is much more easily metabolisable and if it is in the environment, there is no need for those enzymes. On the other hand, the down-regulated genes in Straw 24h belong to GO terms related to ribosomes, transcription and mitochondrial functions. Most of these functions are restored with the addition of glucose. Furthermore, the down-regulation of the mycelium development, amino-acids biosynthesis functions might indicate that the fungus is still not growing after 24h of incubation in wheat straw.

Nitsche *et al.* (2012) observed a down-regulation of the mitochondrial function as part of the transcriptomic response when *A. niger* is exposed to a carbon starvation stress. The response observed in *T. versatilis* could be related to the difficulties to use the complex polymers present in the media, which can be perceived at first as a lack of carbon source, as it has been proposed in the model presented by Delmas *et al.* (2012).

Table 5.3: GO enrichment results for the significantly up or down-regulated genes.

GO ID	GO Term	In Straw 24h vs. Glucose 48h	In Straw+glucose 5h vs. Straw 24h
Biological process			
GO:0045493	xylan catabolic process	Up	Down
GO:0005975	carbohydrate metabolic process	Up	Down
GO:0000272	polysaccharide catabolic process	Up	Down
GO:0045490	pectin catabolic process	Up	Down
GO:0072659	protein localization to plasma membrane	Up	Down
GO:0001676	long-chain fatty acid metabolic process	Up	Down
GO:0051050	positive regulation of transport	Up	Down
GO:0045184	establishment of protein localization	Up	
GO:0009116	nucleoside metabolic process	Up	
GO:0071704	organic substance metabolic process	Up	
GO:0006123	mitochondrial electron transport, cytochrome c to oxygen	Down	Up
GO:0006744	ubiquinone biosynthetic process	Down	Up
GO:0015986	ATP synthesis coupled proton transport	Down	Up
GO:0042254	ribosome biogenesis	Down	Up
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	Down	Up
GO:0006412	translation	Down	Up
GO:0006814	sodium ion transport	Down	Up
GO:0015992	proton transport	Down	Up
GO:0043581	mycelium development	Down	Up
GO:0000447	endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	Down	Up
GO:0009094	L-phenylalanine biosynthetic process	Down	Up
GO:0006364	rRNA processing	Down	Up
GO:0009423	chorismate biosynthetic process	Down	Up
GO:0006571	tyrosine biosynthetic process	Down	Up
GO:0000162	tryptophan biosynthetic process	Down	Up
GO:0006119	oxidative phosphorylation	Down	Up
GO:0009099	valine biosynthetic process	Down	Up

Chapter V

GO:0009098	leucine biosynthetic process	Down	Up
GO:0009073	aromatic amino acid family biosynthetic process	Down	Up
GO:0009097	isoleucine biosynthetic process	Down	Up
GO:0006167	AMP biosynthetic process	Down	Up
GO:0022900	electron transport chain	Down	Up
GO:0006418	tRNA aminoacylation for protein translation	Down	Up
GO:0006561	proline biosynthetic process	Down	
GO:0033615	mitochondrial proton-transporting ATP synthase complex assembly	Down	
GO:0016226	iron-sulfur cluster assembly	Down	
GO:0000002	mitochondrial genome maintenance	Down	
GO:0006414	translational elongation		Up
GO:0006099	tricarboxylic acid cycle		Up
GO:0006415	translational termination		Up
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay		Up
GO:0001731	formation of translation preinitiation complex		Up
GO:0019083	viral transcription		Up
GO:0006522	alanine metabolic process		Up
GO:0006122	mitochondrial electron transport, ubiquinol to cytochrome c		Up
GO:0009085	lysine biosynthetic process		Up
GO:0009088	threonine biosynthetic process		Up
GO:0006536	glutamate metabolic process		Up
GO:0006614	SRP-dependent cotranslational protein targeting to membrane		Up
GO:0006413	translational initiation		Up
GO:0006144	purine nucleobase metabolic process		Up
GO:0006531	aspartate metabolic process		Up
GO:0009060	aerobic respiration		Up
GO:0031120	snRNA pseudouridine synthesis		Up
GO:0006446	regulation of translational initiation		Up
GO:0006574	valine catabolic process		Down
GO:0008152	metabolic process		Down
GO:0006552	leucine catabolic process		Down
GO:0006629	lipid metabolic process		Down
GO:0006550	isoleucine catabolic process		Down
GO:0005982	starch metabolic process		Down
GO:0009251	glucan catabolic process		Down
GO:0044275	cellular carbohydrate catabolic process		Down
GO:0009117	nucleotide metabolic process		Down
GO:0005985	sucrose metabolic process		Down
GO:0018874	benzoate metabolic process		Down
GO:0010411	xyloglucan metabolic process		Down

GO:0030245	cellulose catabolic process		Down
GO:0015672	monovalent inorganic cation transport		Down
GO:0006699	bile acid biosynthetic process		Down
Cellular component			
GO:0005576	extracellular region	Up	Down
GO:0005777	peroxisome	Up	Down
GO:0044459	plasma membrane part	Up	Down
GO:0005840	ribosome	Down	Up
GO:0005759	mitochondrial matrix	Down	Up
GO:0005758	mitochondrial intermembrane space	Down	Up
GO:0005747	mitochondrial respiratory chain complex I	Down	Up
GO:0000276	mitochondrial proton-transporting ATP synthase complex, coupling factor F(o)	Down	Up
GO:0005743	mitochondrial inner membrane	Down	Up
GO:0005739	mitochondrion	Down	Up
GO:0030686	90S preribosome	Down	Up
GO:0045277	respiratory chain complex IV	Down	Up
GO:0030529	ribonucleoprotein complex	Down	Up
GO:0070469	respiratory chain	Down	Up
GO:0005751	mitochondrial respiratory chain complex IV	Down	Up
GO:0032040	small-subunit processome	Down	Up
GO:0005762	mitochondrial large ribosomal subunit	Down	Up
GO:0018444	translation release factor complex	Down	
GO:0044424	intracellular part	Down	
GO:0022625	cytosolic large ribosomal subunit		Up
GO:0033290	eukaryotic 48S preinitiation complex		Up
GO:0015934	large ribosomal subunit		Up
GO:0005763	mitochondrial small ribosomal subunit		Up
GO:0016282	eukaryotic 43S preinitiation complex		Up
GO:0022627	cytosolic small ribosomal subunit		Up
GO:0005852	eukaryotic translation initiation factor 3 complex		Up
GO:0005960	glycine cleavage complex		Up
GO:0005737	cytoplasm		Up
GO:0015935	small ribosomal subunit		Up
GO:0031429	box H/ACA snoRNP complex		Up
GO:0005832	chaperonin-containing T-complex		Up
GO:0005730	nucleolus		Up
GO:0030687	preribosome, large subunit precursor		Up
GO:0070993	translation preinitiation complex		Up
GO:0005835	fatty acid synthase complex		Down

Molecular function			
GO:0031176	endo-1,4-beta-xylanase activity	Up	Down
GO:0030248	cellulose binding	Up	Down
GO:0030246	carbohydrate binding	Up	Down
GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	Up	Down
GO:0008810	cellulase activity	Up	Down
GO:0022857	transmembrane transporter activity	Up	Down
GO:0004567	beta-mannosidase activity	Up	Down
GO:0003824	catalytic activity	Up	Down
GO:0022891	substrate-specific transmembrane transporter activity	Up	Down
GO:0004129	cytochrome-c oxidase activity	Down	Up
GO:0004812	aminoacyl-tRNA ligase activity	Down	Up
GO:0003735	structural constituent of ribosome	Down	Up
GO:0008137	NADH dehydrogenase (ubiquinone) activity	Down	Up
GO:0046933	hydrogen ion transporting ATP synthase activity, rotational mechanism	Down	Up
GO:0003954	NADH dehydrogenase activity	Down	Up
GO:0019843	rRNA binding	Down	Up
GO:0003723	RNA binding	Down	Up
GO:0046961	proton-transporting ATPase activity, rotational mechanism	Down	Up
GO:0051536	iron-sulfur cluster binding	Down	
GO:0004601	peroxidase activity	Down	
GO:0016740	transferase activity	Down	
GO:0016772	transferase activity, transferring phosphorus-containing groups	Down	
GO:0051287	NAD binding		Up
GO:0016597	amino acid binding		Up
GO:0000049	tRNA binding		Up
GO:0050661	NADP binding		Up
GO:0003743	translation initiation factor activity		Up
GO:0004316	3-oxoacyl-[acyl-carrier-protein] reductase (NADPH) activity		Down
GO:0052689	carboxylic ester hydrolase activity		Down
GO:0004560	alpha-L-fucosidase activity		Down
GO:0008171	O-methyltransferase activity		Down
GO:0046556	alpha-N-arabinofuranosidase activity		Down
GO:0008092	cytoskeletal protein binding		Down
GO:0008422	beta-glucosidase activity		Down
GO:0033946	xyloglucan-specific endo-beta-1,4-glucanase activity		Down
GO:0003995	acyl-CoA dehydrogenase activity		Down

5.2. Analysis of the expression of the GH-encoding genes by RT-qPCR

5.2.1. Tracking the best reference genes for RT-qPCR data normalisation

To understand the main motivations of this study, we need to keep in mind that the reverse transcriptase quantitative PCR is a suitable technique to study a limited number of genes in a large variety of conditions. However, data normalisation is a critical step of the workflow that can lead to incorrect biological interpretation of gene regulation if the reference is not validated correctly. In the frame of our work, the transcriptomic analysis of *T. versatilis* exposed to glucose and wheat straw (the details are presented in the previous section) showed that most of the classic reference genes exhibited expression changes in the three conditions tested in that experiment. These findings prompted us to identify a list of putative reference genes. The main highlights of the work are briefly summarized below. This work was the object of a scientific article that was submitted to BMC Genomics. The full article and accompanying additional files can be found in Appendix II and III.

The pre-selection of putative reference genes was done using the RNA-seq data (see section 5.1). The genes selected were stably expressed in the transcriptomic study, had no alternative splicing or antisense transcripts and their expression level was above 30 RPKM. The functional redundancy between the genes selected was avoided in order to minimise the risk of co-regulation. The expression stability of these genes was validated by RT-qPCR in *T. versatilis* cultivated under more than 30 different relevant conditions. The raw Cq values were then used to calculate the relative quantities which were analysed with the geNorm algorithm to rank the candidate genes according to their *M* value and to ascertain their expression stability over all the conditions (Figure 5.11 a). Vandesompele *et al.* (2002) stated that an *M* value below 1.5 indicates good stability. We similarly classified these genes in restricted conditions subsets to identify the best reference genes for normalisation purposes under more focused contexts. The subsets that have been analysed

corresponded to samples obtained from mycelium grown in different carbon sources ('C sources' subset), to samples from mycelium exposed to different stress conditions ('Stress' subset) and to samples harvested during conidial germination ('Germination' subset). We also evaluated the optimal number of reference genes required for accurate normalisation in these sets of conditions (Figure 5.11 b). The remarkable stability for most of the genes, shown by the low *M* value determined by geNorm, which is always below the cutoff value, was also supported by the pairwise variation values that need to be below 0.15 to be acceptable. Then, we probed that no matter which combination of at least 3 of the reference genes selected in this work is suitable for good normalisation. On the contrary, the use of a single reference gene is not recommendable, since an important bias in the transcript levels calculated can be introduced, leading to errors in the interpretation of the data and its biological meaning, in particular when measuring subtle transcriptional changes. This is illustrated in Figure 5.12 b, where the transcript levels of the 2 genes of interest normalised with single genes can lead to under- or over-estimation compared to the normalisation by 3 reference genes.

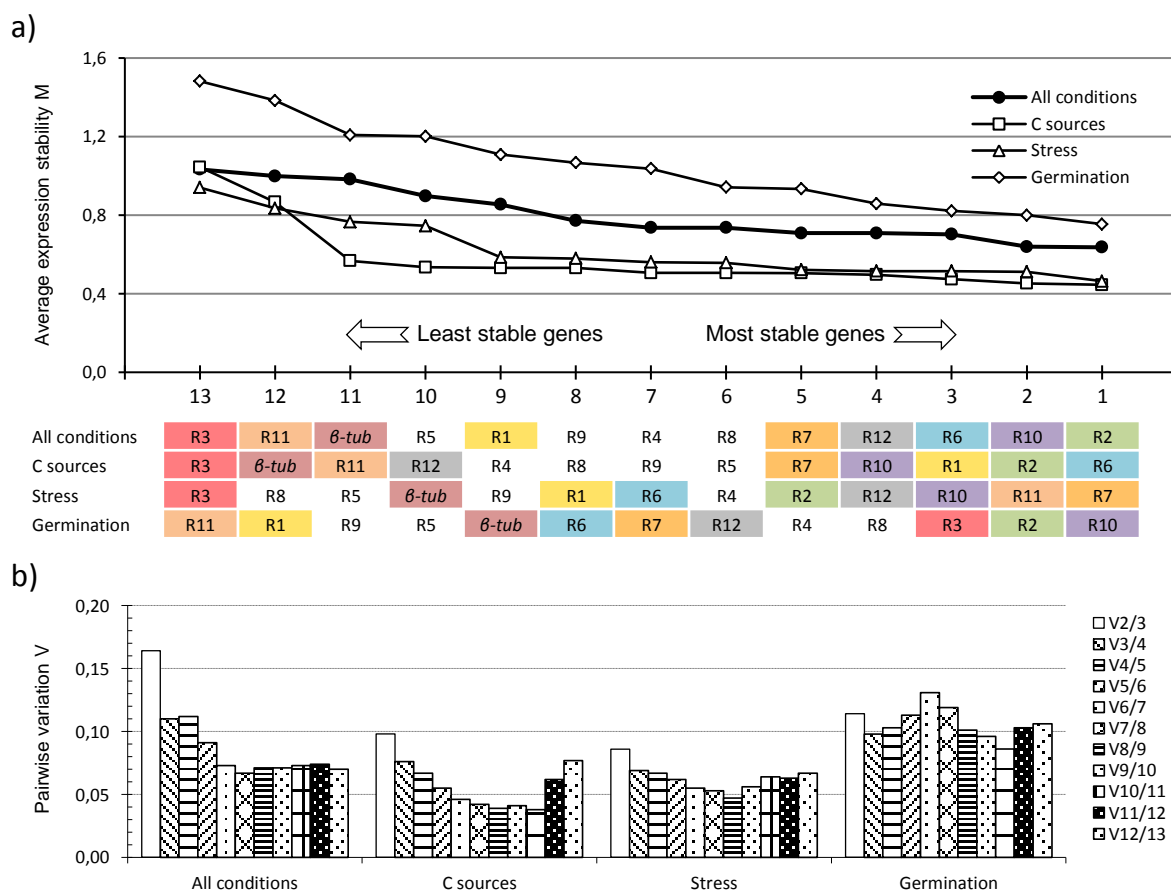


Figure 5.11: geNorm-based ranking of the putative reference genes. a) Genes were ranked from the least stable (on the left) to the most stable (on the right) according to their M value (Y axis). This classification was independently performed by using different sets of conditions: the ‘All conditions’ included the whole set of culture conditions studied by RT-qPCR in this work; the ‘C sources’ subset gathered the 18 samples obtained from growth on different sugars; the ‘Stress’ subset corresponded to 6 samples harvested during stress exposition; the ‘Germination’ subset included the 6 germination time points. For each set of conditions, the result of the classification was given below the X-axis (arbitrary colours attributed to each gene for the sake of clarity). b) Result of the pairwise variation analysis between NF_n and NF_{n+1} to determine the optimal number of genes for reliable normalization. Values below the 0.15 threshold mean that n genes might be sufficient.

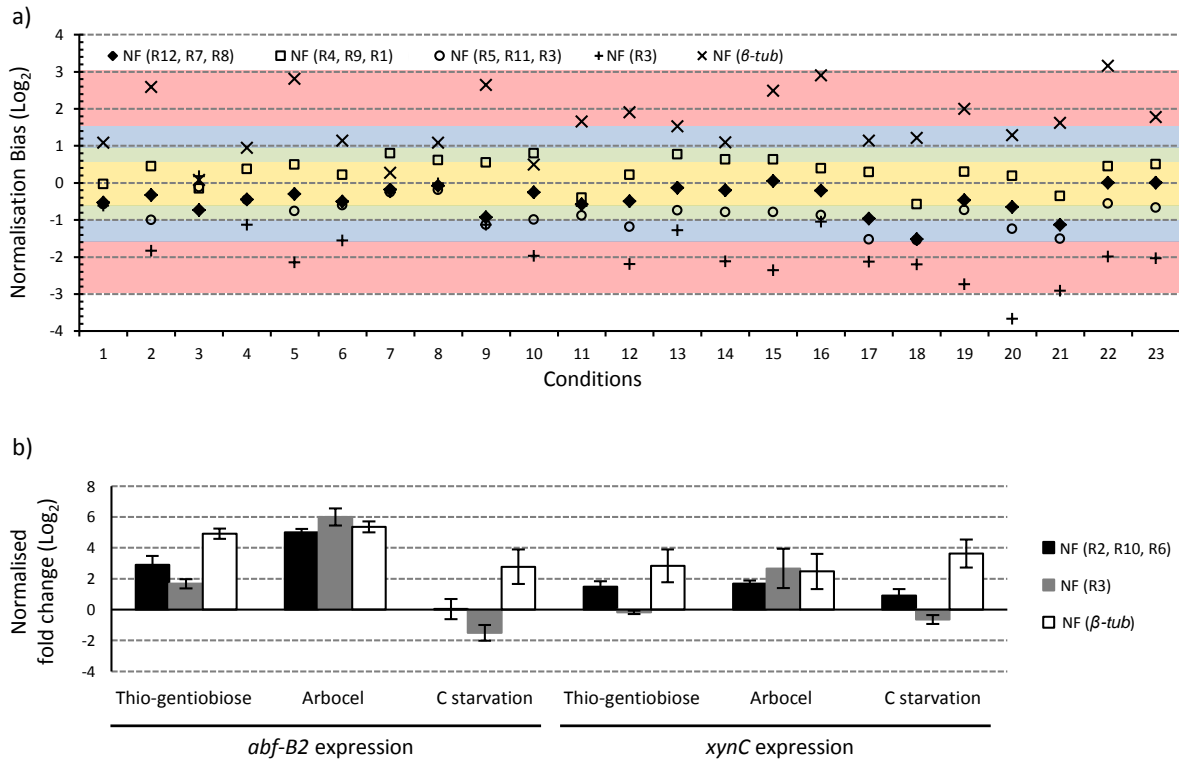


Figure 5.12: Normalization bias analysis according to reference genes selection. For each condition of interest (X axis, e.g. #22 was the thio-gentiobiose sample), we calculated a normalization bias (i.e. under- or over-estimation of the normalised expression value of GOIs) as the ratio between the theoretically best NF ($NF_{(R2, R10, R6)}$ as determined from geNorm classification by using the entire set of conditions) and NF calculated from other combinations of reference genes. $\log_{(base\ 2)}$ of the normalization bias is represented on the Y axis. Yellow zone: less than 1.5 fold bias; Green zone: 1.5–2 fold bias; Blue zone: 2–3 fold bias; Red zone: 3–8 fold bias. a) Comparison of $NF_{(R2, R10, R6)}$ to NF calculated from less and less stable genes, as well as from single genes such as R3 and β -tub. b) Example of quantitative expression results of *abf-B2* and *xynC* under different culture conditions (Thio-gentiobiose, ArboceI and C starvation samples) using 3 different NFs: $NF_{(R2, R10, R6)}$, $NF_{(R3)}$ and $NF_{(\beta-Tub)}$. $\log_{(base\ 2)}$ of FC values on the Y axis (mean \pm SD, n=2 in this experiment).

To demonstrate that some of the candidates suitable for *T. versatilis* can be proposed as promising reference genes for data normalisation in RT-qPCR analysis in filamentous fungi, we collected 92 RNA-seq based transcriptomic datasets from 18 phylogenetically distant filamentous fungi, including industrially important or model species as well as plant or animal interacting fungi.

The expression data of the putative reference genes selected above, but also the “classic” housekeeping genes, like the actin (C1, *act*), the aminopeptidase C (C2, *apsC*), the cytochrome C oxidase subunit V (C3, *coxV*), glyceraldehyde-3-phosphate dehydrogenase, (C4, *gapdh*), etc. and the 4 reference genes identified in *S. cerevisiae* (Teste *et al.*, 2009) were recovered from the RNA-seq datasets. These 100 expression values (log-transformed) were pooled together and analysed for each gene independently to recover two important statistics, *i.e.* the median and the interquartile of the distribution. To search for groups of genes presenting similar behaviour, we conducted a cluster analysis (HAC), using the median and the interquartile as variables. Moderate partitioning led to the identification of 5 classes, highlighted in the interquartile *versus* median scatter plot (Figure 5.13). The best class (red), whose centroid was characterised by a median at zero and the lowest dispersion, contained three of the new candidates (R2 (*ubcB*) and R10 (*sac7*), previously designed as the best reference genes for *T. vesatilis*, and R4 (*fis1*)), but also C10 (*sarA*) that encodes a secretion associated GTP-binding protein and that was already identified as a good reference gene (Bohle *et al.*, 2007; Steiger *et al.*, 2010), and two genes homologous to *S. cerevisiae* *TFC1* (Sc3) and *UBC6* (Sc4).

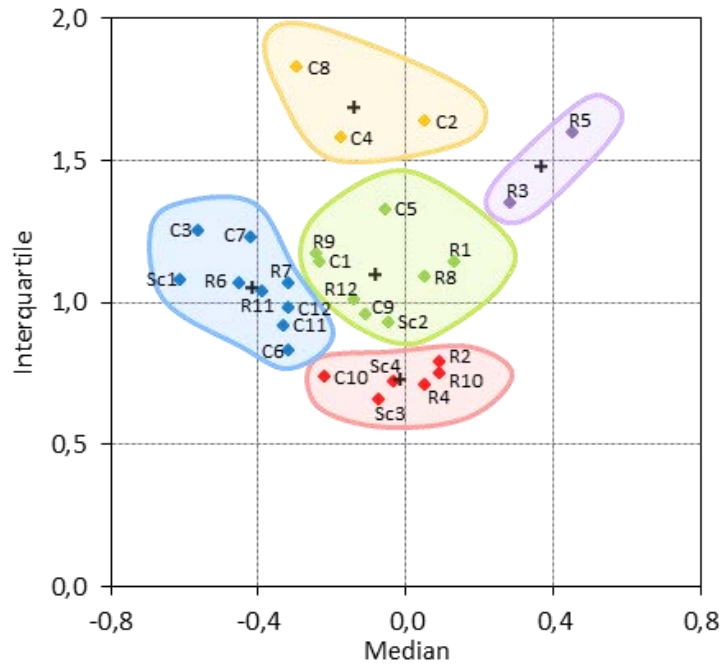


Figure 5.13: Classification of the reference genes according to their median and interquartile. Scatter plot of the interquartile *versus* median. The clusters that were obtained by hierarchical classification (HAC) are circled with different colours. The black crosses indicate the centroid of each cluster.

5.2.2. *RT-qPCR analysis of the expression levels of GH-encoding genes over different carbon sources*

In order to investigate the response of a selection of genes of interest, that are listed on Table 5.4, to different carbon sources or stress conditions (listed on Table 5.5), the mycelium of *T. versatilis* grown on TvMM + glucose was transferred to fresh media containing other carbon sources, or exposed to stress conditions, as described in the materials and methods section (Chapter 2). RNA was extracted and the transcript levels were measured using RT-qPCR. The normalisation of the data was carried out using the R2, R6 and R10 reference genes (for more details on the reference genes, see section 5.2.1). The genes studied include genes coding for enzymes that degrade the cellulose or the hemicelluloses, or even proteins that do not have an enzymatic activity, but that might be

indirectly involved in the degradation of such substrates. In every case, the normalized fold change was calculated using the glucose condition as the control (calibrator sample).

Table 5.4: List of genes of interest studied using RT-qPCR.

Pair of primers	Gene	Annotation
GOI2	Pf200c030g01540	Cellobiohydrolase II -GH6
GOI4	Pf360c030g01900	Cellobiohydrolase I -GH7
GOI6	Pf200c012g01410	Arabinofuranosidase B2 - GH54
GOI7	Pf170c016g00470	Arabinofuranosidase B1 - GH54
GOI10	Pf390c014g00360	Xylanase B - GH11
GOI12	Pf300c053g00050	Xylanase C - GH11
GOI14	Pf340c013g00740	Xylanase D - GH10
GOI15	Pf120c019g00490	β -glucosidase - GH1
GOI19	Pf300c048g00360	β -xylosidase - GH3
GOI23	Pf300c045g00020	β -xylosidase - GH43
GOI26	Pf160c028g00400	Endoarabinase - GH43
GOI27	Pf160c026g01520	α -L-arabinofuranosidase - GH43
GOI30	Pf350c013g00110	GH61
GOI31	Pf120c037g00350	Hydrophobin
GOI33	Pf170c005g00140	Swollenin

Table 5.5: List of carbon sources and stress conditions to which *T. versatilis* has been exposed in order to measure the expression levels of the genes of interest.

	Carbon source/Stress condition	Concentration	Time of exposure
AL37	Xylose	0.2%	30 minutes
AL38	Xylose	0.2%	2 hours
AL39	Xylose	1%	1 hour
AL12	Xylose	1%	24hours
AL40	Arabinose	0.2%	30 minutes
AL41	Arabinose	0.2%	2 hours
AL42	Arabinose	1%	1 hour
AL14	Arabinose	1%	24hours
AL2	Cellobiose	0.2%	30 minutes
AL3	Cellobiose	0.2%	2 hours
AL5	Thio-gentiobiose	0.2%	30 minutes
AL6	Thio-gentiobiose	0.2%	2 hours
AL8	Xylobiose	0.2%	30 minutes
AL9	Xylobiose	0.2%	2 hours
AL10	Xylobiose	0.2%	6 hours
AL16	Avicel	1%	24 hours
AL18	Arbocel	1%	24 hours
AL20	Xylan	1%	24 hours
AL22	Wheat straw	1%	24 hours
AL24	Wheat bran	1%	24 hours
AL25	40°C		1 hour
AL26	pH 2		1 hour
AL27	pH 8		1 hour
AL28	KCl	0.5 M	1 hour
AL29	N starvation		1 hour
AL30	C starvation		1 hour

5.2.2.1. *Global analysis of the expression patterns of the hydrolytic enzymes*

In Figure 5.14 the box-plot represents the distribution of the expression values (in a logarithmic scale) that were obtained over the 25 conditions, for each gene. The genes have been sorted by the size of the interquartile, which contains the central 50% of the values. In this way, it is possible to see that there are genes that have a larger variability than others. The median, represented by the dark line within the box, indicates whether the GOIs were globally up- or down-regulated. For instance, the *xynC* (GOI12), the *abf-B1* (GOI7), the arabinofuranosidase GH43 (GOI27) and the GH61 (GOI30) genes not only have reduced variability, but also they were not prone to huge activation in response to changing conditions, with median FC close to 1 and most of the points that do not exceed 4-fold change (above the limit of the interquartile). The hydrophobin (GOI31) gene, and to a lesser extent, the swollenin (GOI33) gene, are the only genes of the list that were globally down regulated under the conditions of interest. Only a few conditions led to extreme values that outburst with significant over-expression. On the other hand, there are other genes that show a wide variation in their response to changing environment, such as the β -glucosidase (GOI15), the endoarabinase (GOI26), the *abf-B2* (GOI6), the *xynB* (GOI10) and *xynD* (GOI14), the *cbhII* (GOI2) and *cbhI* (GOI4) or the β -xylosidase (GOI19, GOI23) genes. The most responsive gene was GOI14 that showed almost 2,000-fold induction for the upper limit of the box. It is interesting to notice that genes that encode similar proteins (highlighted by the color code of the boxes) presented very different behaviour when looking at the data in general.

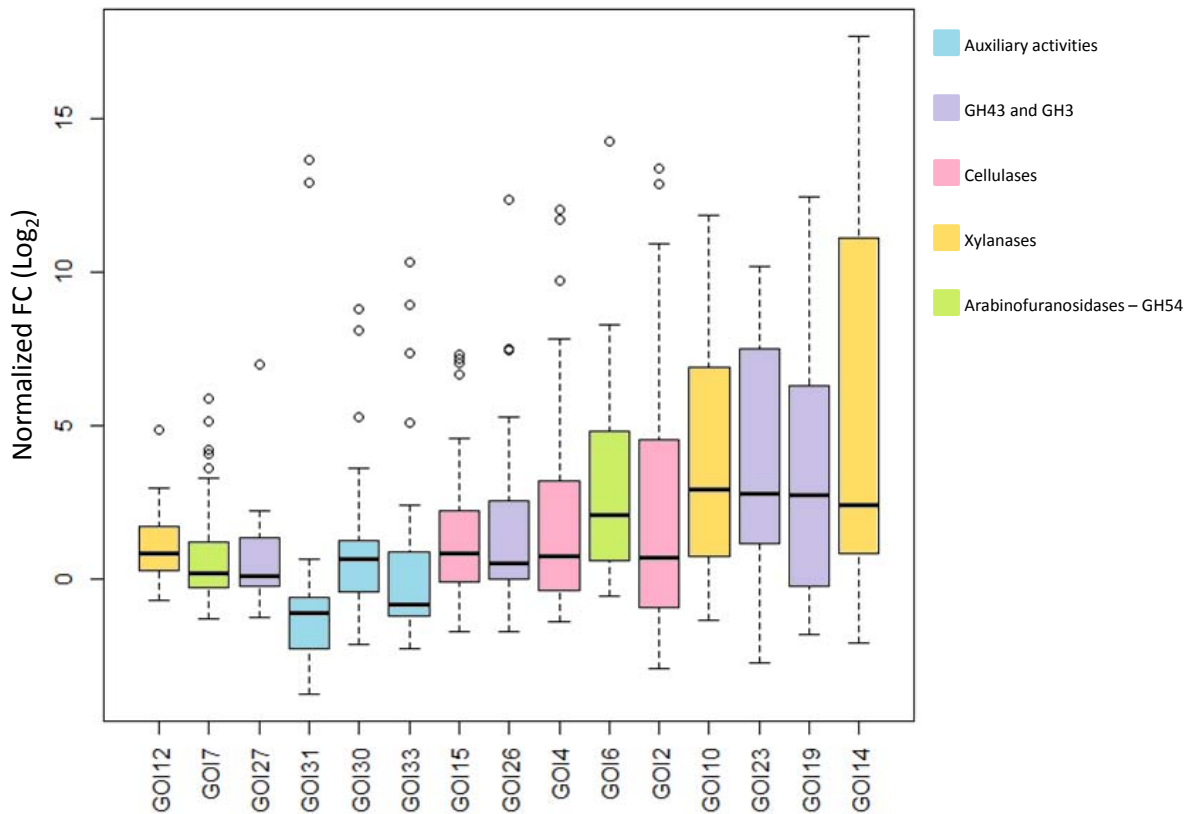


Figure 5.14: Distribution of the expression values for each gene in all conditions, expressed in logarithm. The lower and upper boundaries of the box represent the 25th and the 75th percentile, respectively. The line within the box represents the median and the whiskers indicate the highest and the lowest data. The outliers are represented by the circles outside the whiskers. The colour of the box indicates the kind of enzyme encoded by the gene.

In order to do a global analysis of the expression of the GOIs, a hierarchical classification using the Ward method was carried out over the data to group the genes according to their transcript levels in all the conditions tested. The result is shown in Figure 5.15. Three distinct classes could be defined, each composed of genes encoding different kinds of enzymes. In the first class the two β -xylosidase genes are clustered very close together (GOI19 and GOI23), and in the same group we find the *abf-B2* (GOI6), the *xynB* (GOI10) and, a little further away, the *xynD* (GOI14). The other two classes are more closely related. One of them is composed by the two cellobiohydrolase genes (GOI1 and GOI1), which cluster together with the endoarabinase (GOI26), the swollenin (GOI33), the β -

glucosidase (GOI15) and the GH61 (GOI30) genes. And the last class is formed by the hydrophobin (GOI31), plus the *abf-B1* (GOI7), the *xynC* (GOI12) and the arabinofuranosidase GH43 (GOI27) genes. It is interesting to note that the HAC classified the genes almost identically than in the boxplot showed above (Figure 5.14); the genes with the lowest expression values and lowest variability are classed together, whereas the genes with high transcript level values and high variability are in another class. But there is no information on the co-regulation of the genes because this aspect is not taken into account by the classification method.

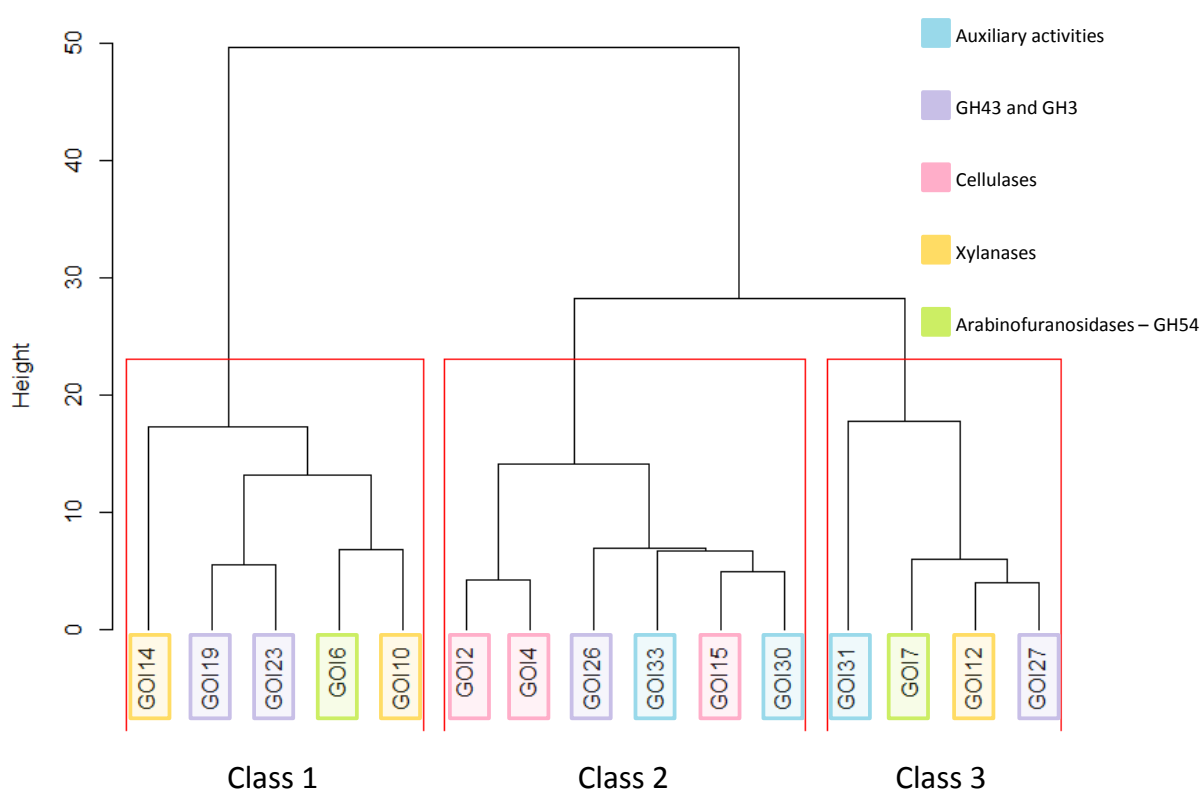


Figure 5.15: Dendrogram of the hierarchical clustering of the genes of interest. The Y axis indicated the distance between the genes. The three classes defined are indicated by the red boxes. The colour boxes containing the gene names indicate the kind of enzymes encoded by the genes.

Conversely, the expression levels obtained for all the genes in each one of the conditions were analysed separately. The distribution of the transcript levels of all the genes are represented in the

boxplot in Figure 5.16. The conditions are sorted according to the size of the box containing the central 50% of the values. In this boxplot, AL1 which corresponds to the Glucose 1% condition is represented by a line, there is no variability for this condition since it is the control condition and the fold change is always set to 1. Interestingly, the stress conditions (AL25 – AL30) are the ones that show the smallest variability and globally lead to slight down-regulation of the expression of the genes. Also, the exposure to arabinose at high concentration (1% - AL42) show low transcript levels for all the GOI in the short term, but the variability increases if the incubation is longer (AL14). The same sugar, but in a low concentration (0.2% - AL40 and AL41) induce a higher fold-change in the transcript level of the genes studied. Interestingly, the polysaccharides (AL16, AL18, AL20, AL22 and AL24), the disaccharides (AL2, AL3, AL5, AL6, AL8, AL9, AL10) and the xylose at 0.2% (AL37, AL38) are the conditions that have the highest variation in the fold-change values. Unquestionably, wheat straw (AL22) is the substrate that induces the highest fold-change of the group of genes under study.

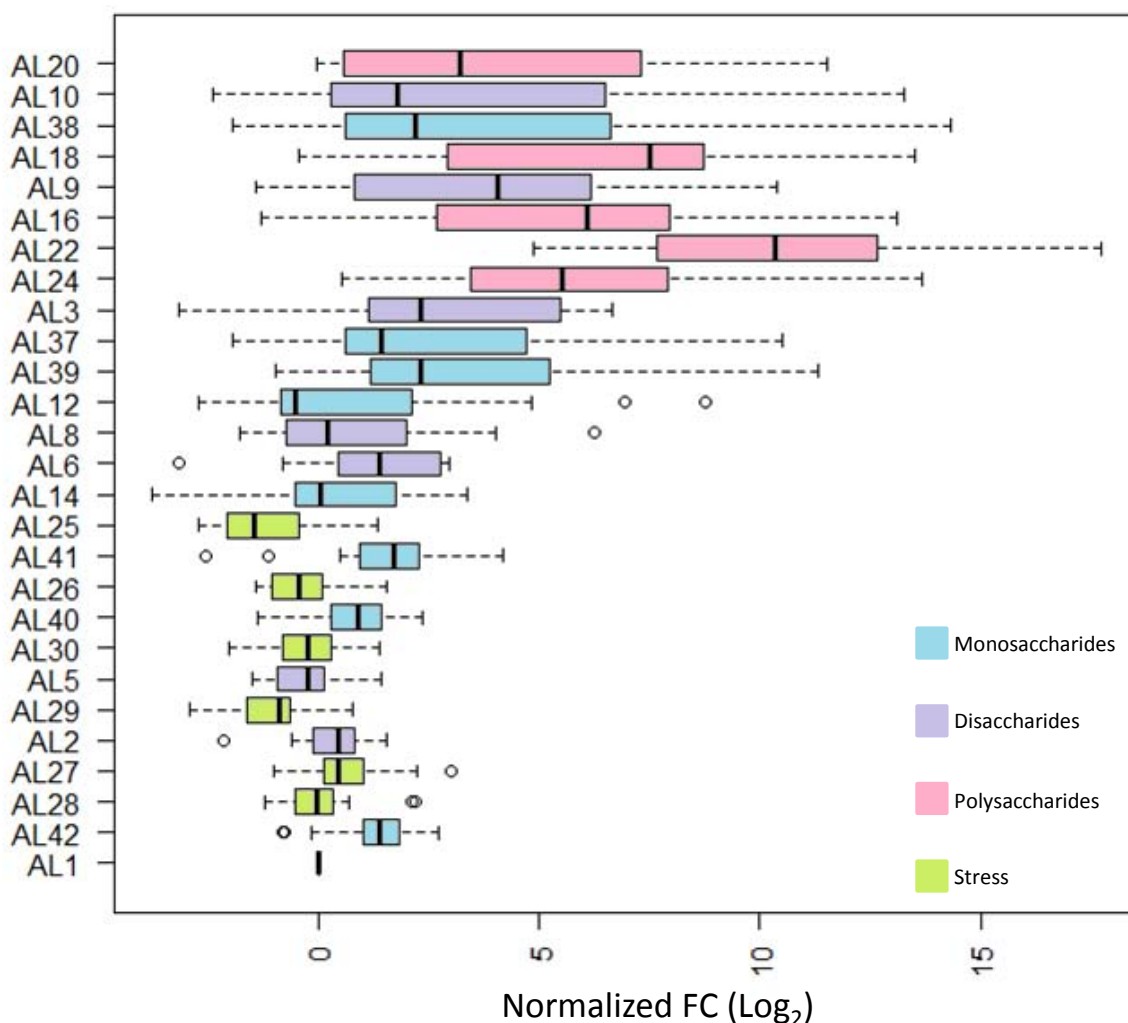


Figure 5.16: Distribution of the expression values for all the genes in each condition, expressed in logarithm. The left and right boundaries of the box represent the 25th and the 75th percentile, respectively. The line within the box represents the median and the whiskers indicate the highest and the lowest data. The outliers are represented by the circles outside the whiskers. The colour of the boxes indicates the kind of condition.

As performed previously to classify the genes as a function of their global behaviour considering all the conditions, a hierarchical classification using the Ward method was performed to group the conditions according to the general response they induce over all the genes studied. The dendrogram shown in Figure 5.17 is the result of this classification. Two very distinct groups can be distinguished: to the left all the conditions that induce rather low fold-change values for all the genes; to the right, the conditions that induce big transcription changes. Five classes were defined: two on the left

branch and three on the right branch. From the left to the right, the class A groups all the stress conditions, the cellobiose 30 minutes and thio-gentiobiose 30 minutes conditions. All these conditions produce very small changes in the transcript levels of the hydrolytic genes studied, and some of them are repressive conditions. The second one, class B, contains all the arabinose-containing conditions, cellobiose 2 hours, thio-gentiobiose 2 hours, xylobiose 30 minutes and xylose 1% 24 hours. These are conditions that produce a slightly higher induction of the genes than the first group, but the two classes are closely related. The third group, class C, includes the xylose, xylobiose and xylan-containing conditions. The response produced by the xylose-bearing compounds is then clearly separated from the other substrates. The last two classes, D and E, concern the polysaccharides; the first one groups the different forms of cellulose, Avicel and Arbocel, and the last one, the two most complex lignocellulosic materials tested, wheat straw and wheat bran. These classes are closely related, but still different from each other. The response induced by the lignocellulosic materials is much more complex than the one induced by the cellulose.

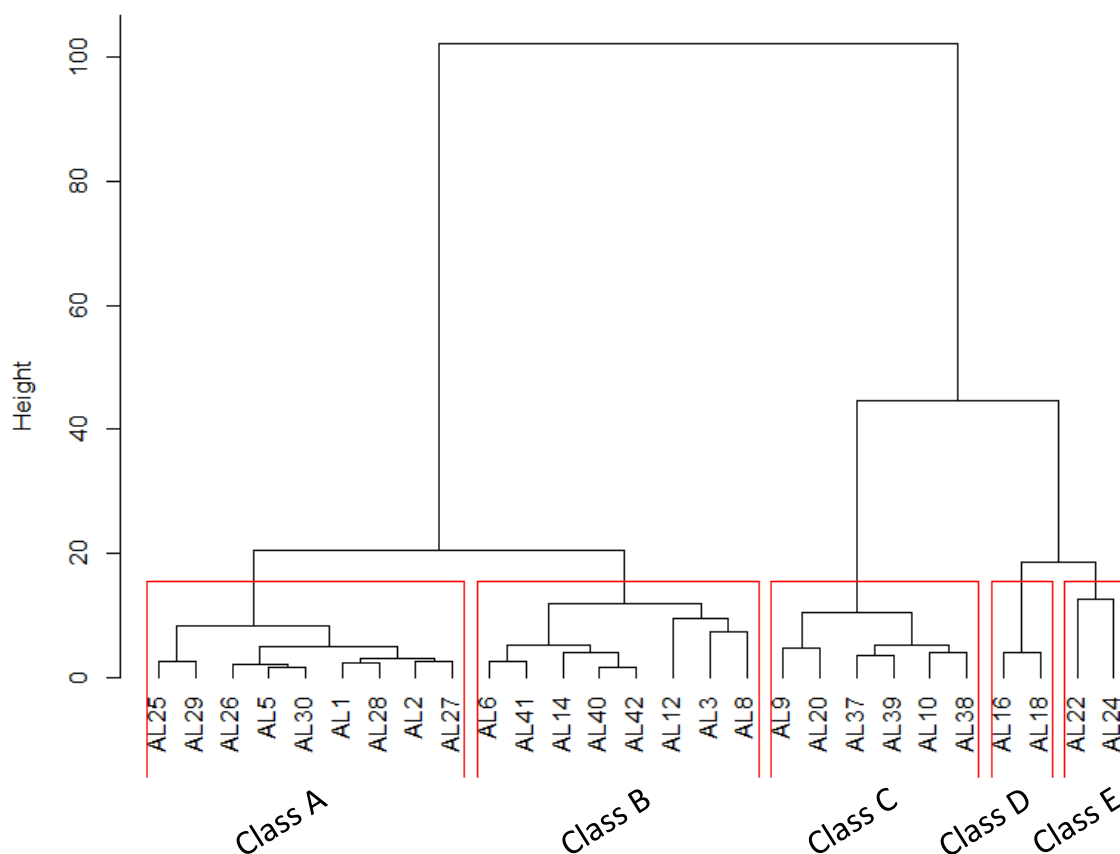


Figure 5.17: Dendrogram of the hierarchical clustering of the conditions. The Y axis indicated the distance between them. The five classes defined are indicated by the red boxes.

The heat map in Figure 5.18 shows the result of a double hierarchical classification of the expression fold-changes, to keep the information on both, the genes and the conditions. This classification allowed identifying sections, indicated with the black lines. In this way, we can observe that the genes in class 3 very small fold change variations in most conditions, except for GOI31, which is strongly induced by the class E (lignocellulose). The genes in class 2 are induced by the cellulose and the lignocellulose from classes D and E, respectively, but they do not show an important change in the transcript level with any of the other substrates or conditions. Finally, the class 1 of genes is highly expressed in the classes C, D and E of conditions; they are also induced by the class B to a lesser extent, but not by class A.

Very few are the cases where the genes are repressed, but the hydrophobin (GOI31) is an example; this gene is repressed or remains stable in most of the conditions, except the lignocellulosic materials. This may indicate that a compound within the lignocellulose is necessary for a full induction of this gene. Further experiments are necessary to determine the regulatory mechanism that controls this gene. The swollenin (GOI33) shows a similar pattern, although it is classified in another group. The stress conditions, grouped in class A, produced a down-regulation of most of the genes of interest, suggesting that the production of hydrolytic enzymes in this kind of situations is not a priority for the fungus. The slight induction observed in some cases may be simply a de-repression, but no full induction could be detected.

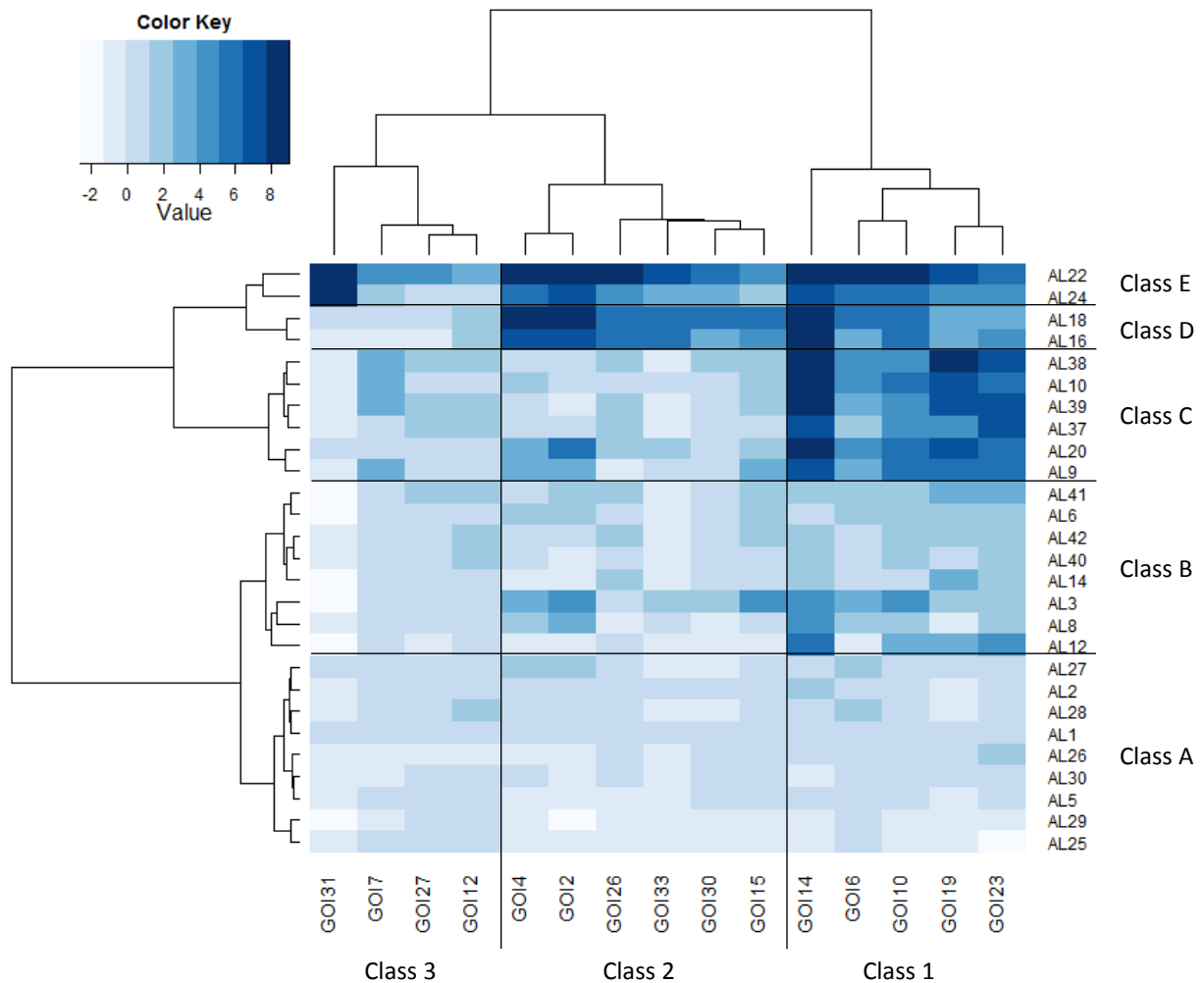


Figure 5.18: Heat map of the hierarchical clustering of the genes of interest, shown in the X axis, versus the conditions, shown in the Y axis. The expression level values are

represented in a logarithmic scale. Light blue represents a low expression value, dark blue represents a high expression value. The colour scale is shown in the upper left corner of the figure.

In order to get a clear idea of the possible co-regulation events, in other words, to identify the genes that behave in the same way throughout the conditions, we used a correlation matrix. Each cell of the matrix is a scatter plot representing the FC values obtained from the 25 conditions, for two GOIs that are compared (see examples in Figure 5.20). The correlation coefficients from each individual scatter plot are summarized in a heat-map (Figure 5.19). Interestingly, the correlation between most of the genes is significantly high. If we consider that a correlation coefficient (r) of 0.38 is statistically significant, most of the genes show a rather good correlation between them, except for the two β -xylosidase genes (GOI19 and GOI23). However, these two genes show a correlation between themselves. Also, the gene encoding for the hydrophobin (GOI31) has a low level of correlation with all the other genes studied, which suggests that this gene has its own regulatory mechanism, independent from the other GH-encoding genes. The correlation of the fold-change values of all the remaining genes may suggest that the regulation systems that control their transcription are somehow related, moreover, some of the genes may share the exact same regulatory systems with much higher correlation coefficients (0.7 or above). Some groups of genes with similar behaviour can be distinguished, like GOI2 with GOI4, GOI30 and GOI33, or the genes GOI6 with GOI10 and GOI14; therefore we decided to group the genes according to the level of correlation between them and also by their biological function and significance. Hence, we could identify 5 different classes. The first one contains the genes coding for proteins involved in the degradation of the cellulose: GOI2, GOI4, GOI15, GOI30 and GOI33. The second class puts together the hemicellulases encoding genes: GOI6, GOI10, GOI14 and GOI26. The third class includes other hemicellulases encoding genes that show a low transcript level in most of the conditions tested: GOI7, GOI12 and GOI27. Class four is only composed by the GOI19 and GOI23, the genes that show a

dissimilar performance to all the other genes. And finally, the GOI31 is isolated from all the other genes and constitutes the fifth class.

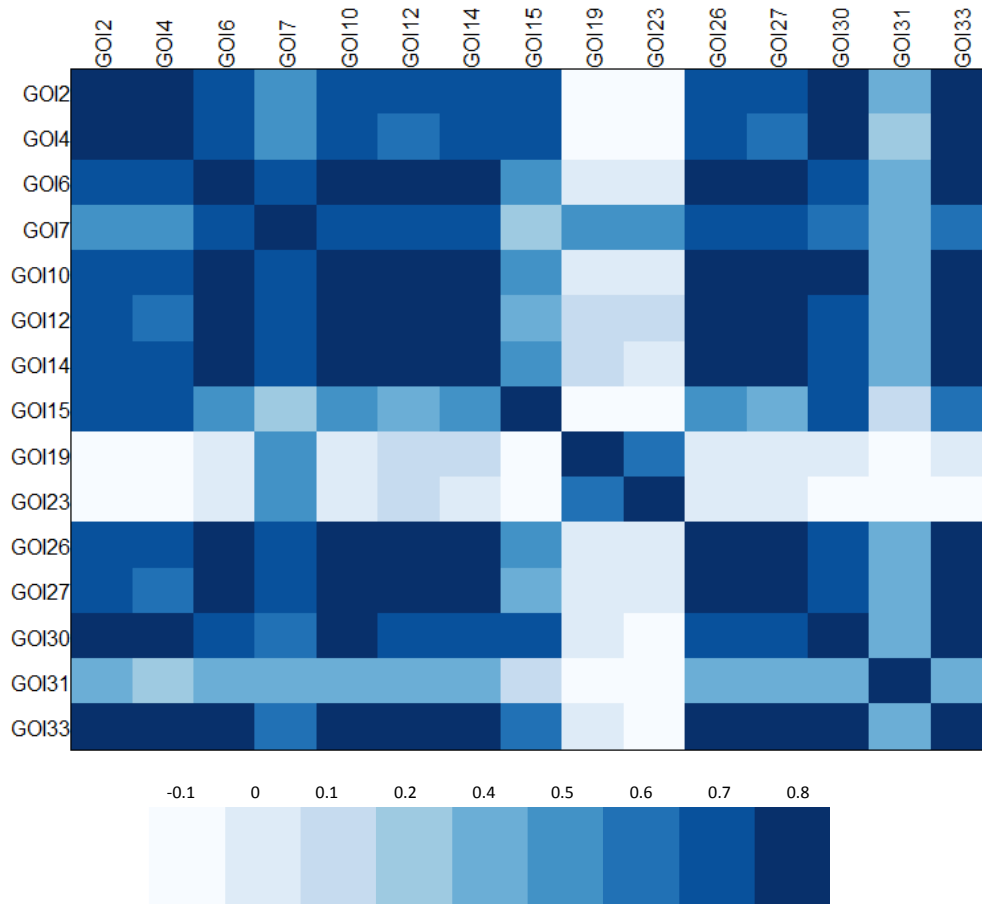


Figure 5.19: Heat map of the correlation between the 15 genes studied, evaluated in 25 different conditions. Each column and each line represent a gene of interest. The low r values, which mean low correlation, are represented by the light blue and r values close to 1 are represented by the dark blue. Scale at the bottom of the figure.

The matrixes of scatter plots done over the genes that belong to the 4 different classes are shown on Figure 5.20. The fifth class concerns only one gene; hence it is not represented in this figure. The correlation within the classes is remarkable, except for class 3 that has the lowest correlation coefficients, yet statistically significant. The genes in this class correspond to the hemicellulases that have a fairly low transcript level in most conditions. However, they have a few points where the transcript levels increase importantly, corresponding to the complex lignocellulosic

material conditions, which appear in the upper, right corner of the scatter plots. In classes 1 and 2 the genes show very good association between them. The two β -xylosidase genes, GOI19 and GOI23, which were clearly less related to the other genes studied, are also very well correlated between them, with only a few outliers that are detached from the majority of the points that are positioned on the diagonal.

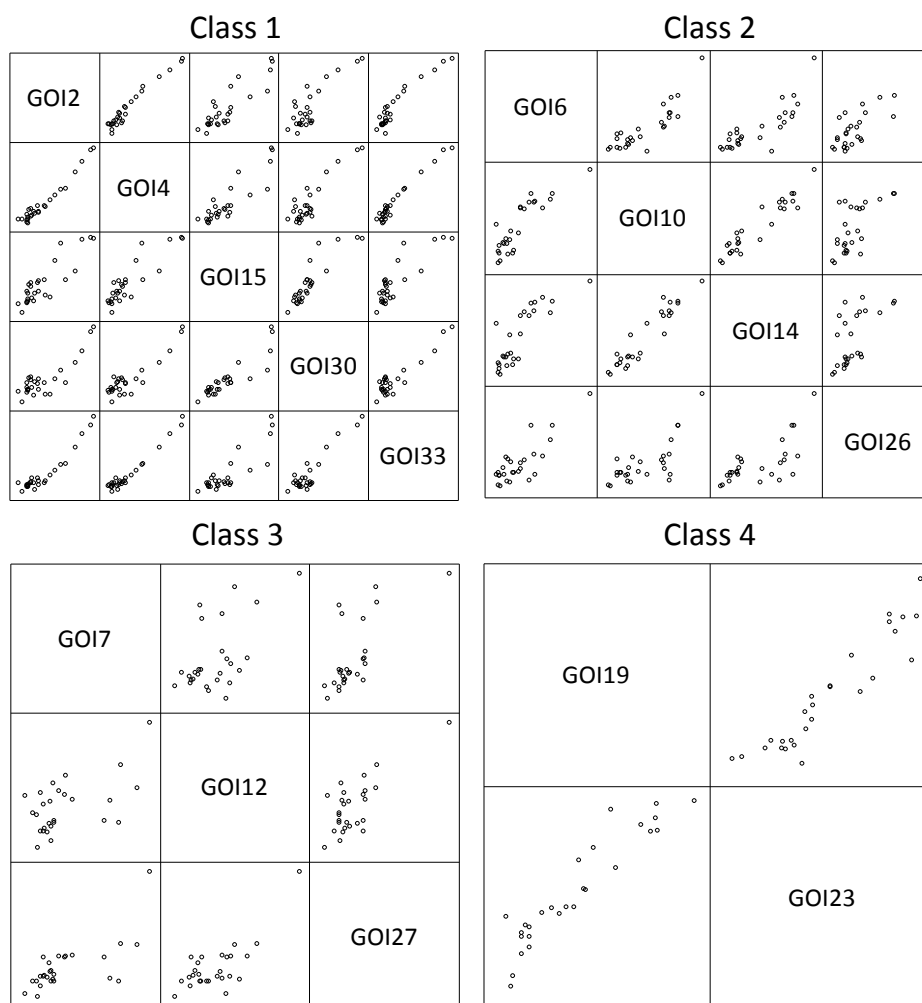


Figure 5.20: Matrix of scatter plot showing the expression levels of the GOI, in a logarithmic scale. Each matrix contains the genes belonging to a class defined by the correlation level. The expression levels of each gene are plotted against those of the other genes.

5.2.2.2. *The class 1 expression pattern: The cellulase genes*

The class 1 is composed by the two cellobiohydrolases (GOI2 and GOI4), the β -glucosidase (GOI15), the GH61 (GOI30) and the swollenin (GOI33) genes. All of the proteins in this class are somehow involved in the degradation of the cellulose.

Most of these genes are moderately induced by the presence of xylose or arabinose in the medium, with an activation of no more than 4-fold. In spite of the induction effect observed after 1 hour of exposure to high concentrations of either xylose or arabinose, a repressive effect in the long term for all the cellulase genes studied was observed (Figure 5.21 a). The high concentration of these sugars could induce a response similar to the carbon catabolite repression observed when glucose is present in the medium, leading to a repression of the hydrolytic enzyme encoding genes. Interestingly, the swollenin is repressed by both monosaccharides, no matter which concentration is used, suggesting that this gene is not under the control of the main activators XlnR or AraR.

All the genes were induced by the disaccharides tested, and the highest peak of induction was observed after 2 hours of exposure to the substrate (Figure 5.21 b). Cellobiose and xylobiose provoked the highest response in these genes, with fold-changes that reach up to 100. The thio-gentiobiose, on the other hand, only induced a response after 2 hours of exposure, and it was the smallest induction observed, in comparison with the other 2 disaccharides. The swollenin, along with the GH61 genes, present a moderate response to the disaccharides, between 0.5 and 5-fold, whereas the 3 other cellulase genes show a wider response that goes from 2 to 100-fold induction.

After 24 hours of incubation with the different polysaccharides, a strong induction was observed for all the genes in all the substrates, the swollenin gene included (Figure 5.21 c). The complex lignocellulosic substrate, wheat straw, was the best inducer, in particular for the two cellobiohydrolase I and II genes, that reached a fold-change of 10,000 (GOI2) and 4,000 (GOI4), respectively. The cellulose, Avicel, and hemicelluloses mixture, Arbocel, also produced a similar response, leading to inductions of up to 7,000-fold. The strong induction observed may suggest the

intervention of other activators besides the CCR release, such as ClrA. Predictably, xylan was the weakest inducer of this group of substrates, with fold-change values between 2 and 160. This response is probably linked to the release of the CCR and the induction through the xylanases activator XlnR.

The high temperature, along with the acid pH, and the lack of nitrogen, produced between 2 and 10-fold repression of the genes in this class. On the other hand, the basic pH induced 2 to 8-fold the expression of the cellobiohydrolase, and the β -glucosidase genes. The carbon starvation produced a < 5-fold down-regulation of all the genes, but a small induction of the β -glucosidase was observed (Figure 5.21 d). This induction may be part of the first response of the fungus to the starvation condition, with the release of the CCR, as it was proposed by Delmas *et al.* (2012). The lack of induction in nitrogen starvation condition indicates that the AreA regulator in *T. versatilis* has different targets than the ones described by Aro *et al.* (2005) in *A. nidulans*, where this regulator activates the cellulase genes in response to the lack of nitrogen.

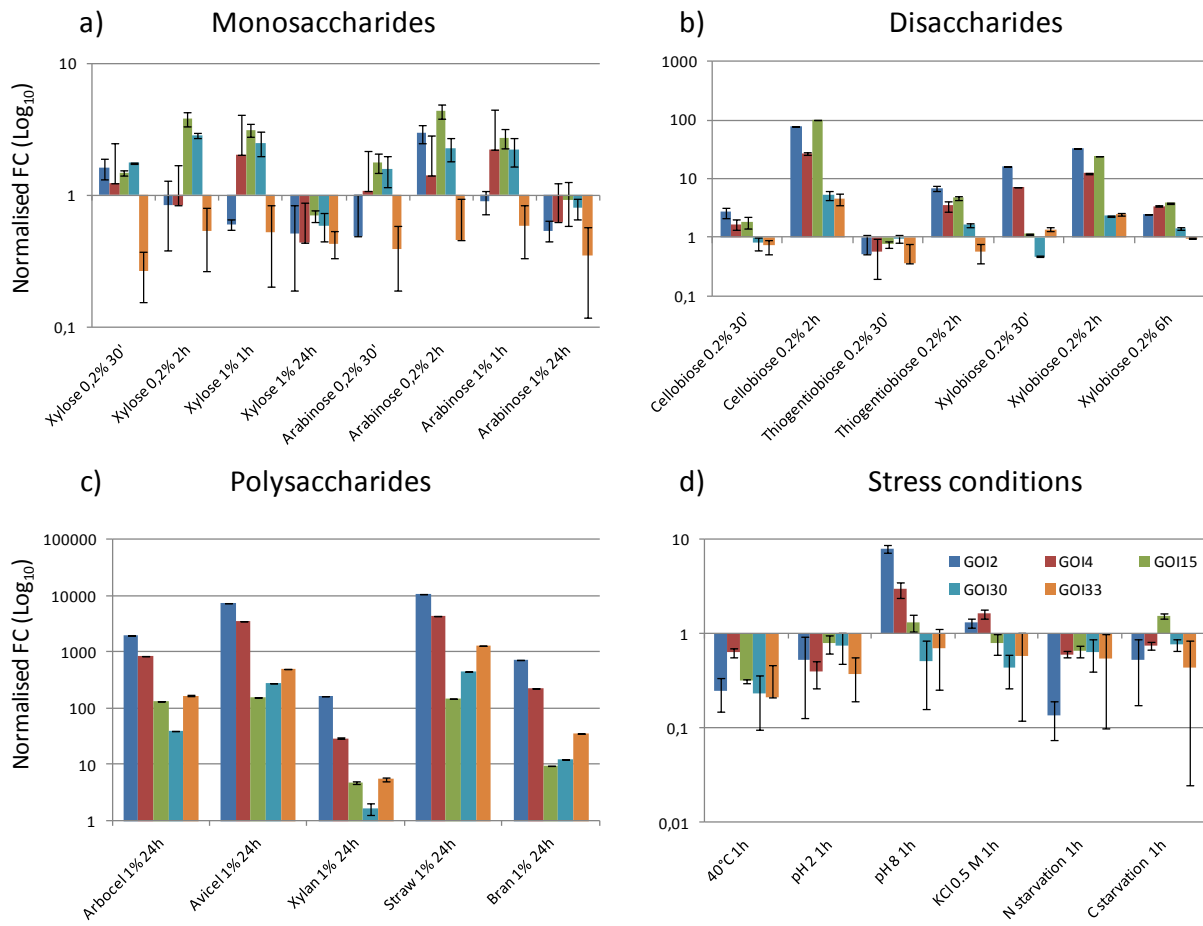


Figure 5.21: qPCR results for the class 1 genes. The *cbhII* is represented by the blue bars, *cbhI* by the red bars, β -glucosidase by the green bars, GH61 by the light blue bars and swollenin with orange bars. The response of the 6 genes was analysed when the fungus is exposed to monosaccharides (a), disaccharides (b), polysaccharides (c) and stress conditions (d) with different concentrations and time of exposure. The fold change (FC) is calculated considering the glucose condition as a reference. The expression values are represented in a logarithmic scale. Error bars represent SEM.

5.2.2.3. The class 2 expression pattern: The main hemicellulase genes

The second class groups the *xynB* (GOI10), *xynD* (GOI14), the *abf-B2* (GOI6) and the endoarabinase (GOI26) genes. The first three of them are involved in the degradation of the arabinoxylan, a kind of hemicellulose. The endoarabinase, on the other hand, is a pectinase involved in the degradation of the arabinan chain.

The two monomers tested, xylan and arabinose, were chosen for this experiment because they are the degradation products of the arabinoxylan and they are supposed to be involved in the induction of the xylanase and cellulase genes through the intermediate action of transcription factors such as XlnR or AraR (Noguchi *et al.*, 2011; de Souza *et al.*, 2013). When the fungus is exposed to xylose or arabinose, all these genes are induced, (Figure 5.22 a). In fact, the induction by xylose is surprisingly high, the fold-change values reach 20,000, and even the high concentration of xylose (1%) produces an important up-regulation of this group. Moreover, the induction is progressive in time when xylose is present in a small concentration (0.2%). The induction observed with arabinose is lower than with xylose, no more than 8-fold, but at low concentration of the substrate the induction is also progressive in time. This suggests that the expression of the hemicellulases encoding genes may be controlled by both XlnR and AraR, but XlnR is the responsible for their full induction. Interestingly, the *xynD* gene shows an enhancement of its expression far more important than the other genes in the group when xylose is present, suggesting that the promoter of this gene is more active than the others and easily inducible with the right substrate.

The three hemicellulase genes in this class were induced by all the disaccharides tested, being the xylobiose the best inducer in this case, which produced up to 9,000-fold induction (Figure 5.22 b). The induction observed with xylobiose reaches the maximum at 6 hours of exposure, but after only 30 minutes there is already a response (from 3 to 76-fold induction). The *xynD* is the one that has the higher transcript levels in these conditions, but it was also the gene that showed the highest transcript levels on xylose. The thio-gentiobiose, which is a sulphur containing analogue of the gentiobiose, a disaccharide containing 2 units of glucose with a β -1,6 linkage, is not as easily metabolized because it cannot be easily degraded by the glucosidases. The cellobiose and xylobiose are rapidly hydrolysed and converted to monomers, but this process is much slow with thio-gentiobiose (Kurasawa *et al.*, 1992). The results obtained with the thio-gentiobiose were not as good as we expected. In fact, this was the weakest inducer tested; it produced no more than 7-fold up-

regulation of the expression of the genes. The cellobiose, on the other hand, produced an induction of up to 100-fold of all the genes, in particular the *xynB* and the *xynD*. The endoarabinase, conversely, did not show any induction in presence of the disaccharides tested, and in some cases this gene was repressed.

The polysaccharides contributed to a high induction of this group of genes; in fact, wheat straw produced the biggest transcriptional activation for all the genes, between 4,000 and 200,000-fold induction (Figure 5.22 c). The cellulose or the xylan pure produced a similar response in the *abf-B2* and the *xynB* genes. The *xynD* gene was, once again, the gene that showed the highest induction compared to the other 3 genes in this class.

Concerning the stress conditions (Figure 5.22 d), the high temperature produced up to 5-fold repression of these genes. The alkaline pH did induce the *abf-B2* (4.8 fold-change) and *xynD* (2.5 fold-change), suggesting that these genes may be controlled by the Pal pathway, although the induction is quite low. The presence of salt in the media did not affect the expression of the genes in this class, except the *abf-B2* which was 4-fold induced. The nitrogen starvation repressed all the expression of the genes about 3-fold. And the carbon starvation only activated the expression of the *xynB* gene 2-fold.

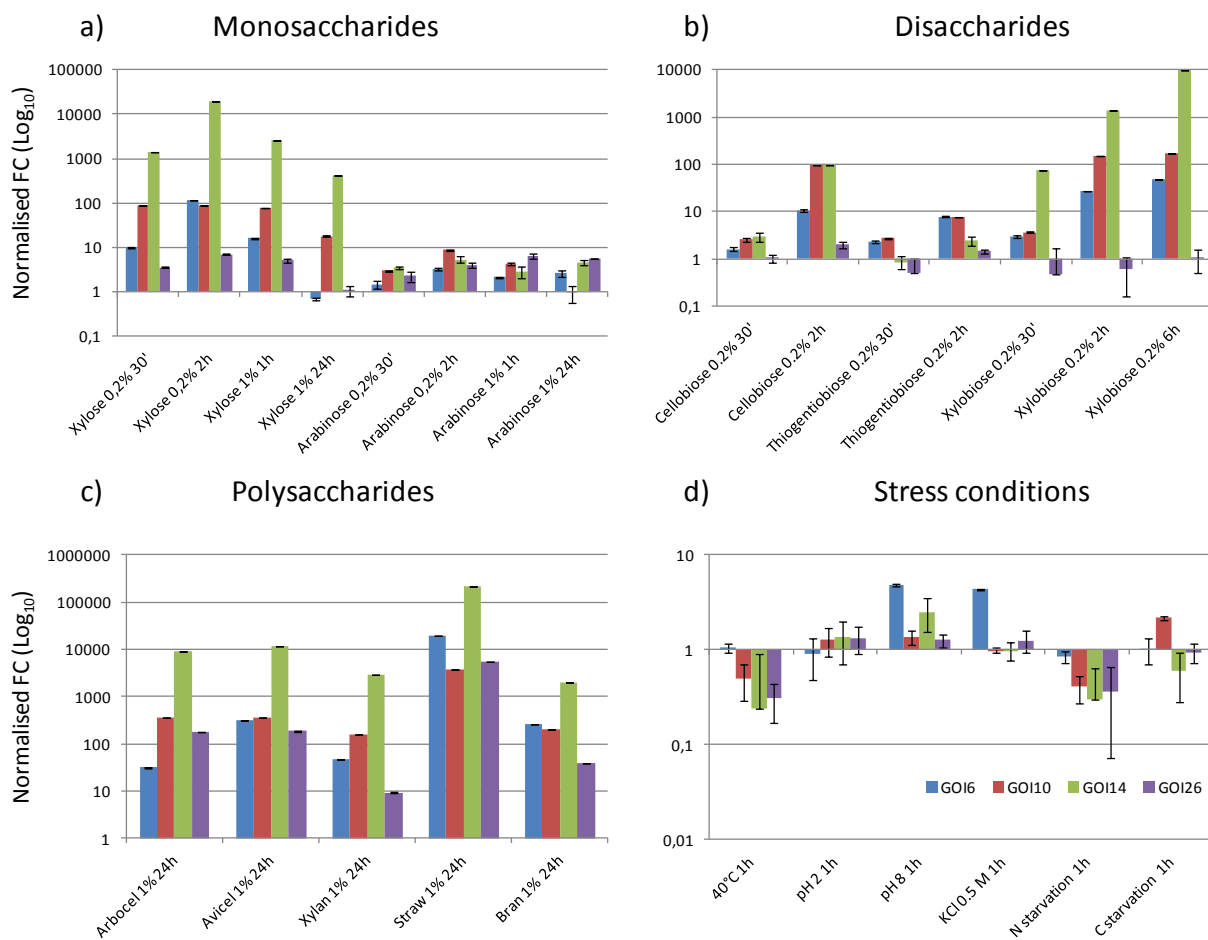


Figure 5.22: qPCR results for the class 2 genes. The *abf-B2* is represented by the blue bars, *xynB* by the red bars, *xynD* by the green bars and the endoarabinase by the violet bars. The response of the genes was analysed when the fungus is exposed to monosaccharides (a), disaccharides (b), polysaccharides (c) and stress conditions (d) with different concentrations and time of exposure. The fold change (FC) is calculated considering the glucose condition as a reference. The expression values are represented in a logarithmic scale. Error bars represent SEM.

5.2.2.4. The class 3 expression pattern: the minor hemicellulase genes

The third class groups the hemicellulase genes that have the lowest transcript levels in most conditions. It includes the *Abf-B1* (GOI7), the *XynC* (GOI12) and the arabinofuranosidase GH43 (GOI27) encoding genes. These enzymes are involved in the deconstruction of the hemicellulose, but they present an expression pattern very different from the other hemicellulase genes studied, suggesting that they are governed by a distinct regulatory mechanism.

The three hemicellulase genes presented 3 to 35-fold induction on xylose, which is very low, compared to the 20,000-fold induction of class 2 hemicellulase genes. A high concentration of xylose completely abolishes the induction of these genes after 24 hours of exposure. The presence of arabinose in the medium produces no more than 6-fold up-regulation of these genes (Figure 5.23 a). When the fungus is exposed to disaccharides, the arabinofuranosidase GH43 genes did not respond to these substrates. Only the *abf-B1* and the *xynC* were induced by the presence of xylobiose, but they only showed 16 and 2-fold induction, respectively (Figures 5.23 b).

Neither the two different kinds of cellulose used, nor the xylan induced an important response of any of the genes in this class, except for a 3-fold up-regulation of the *xynC* gene by the two kinds of cellulose used. Only the complex lignocellulosic materials provoked the 3 genes to be up-regulated from 2 to 130-fold, and wheat straw was the most efficient inducer (Figure 5.23 c).

As to the stress conditions, the high temperature produced a 3-fold induction of the arabinofuranosidase GH43 gene, but the rest of the genes were stably expressed. The acid pH repressed all the genes (2 to 5-fold repression), whereas the alkaline pH did not provoke any significant change in the expression of these genes. The presence of salt stress and the lack of a nitrogen source in the media, as well as the carbon starvation condition induced the expression of the xylanase C gene between 2 and 5-fold (Figure 5.23 d). It is interesting to notice that the xylanase C seems to be a gene that is induced in stress conditions, but none of the molecules that cause the full induction of the expression of the other xylanase genes seem to affect significantly the expression of this gene.

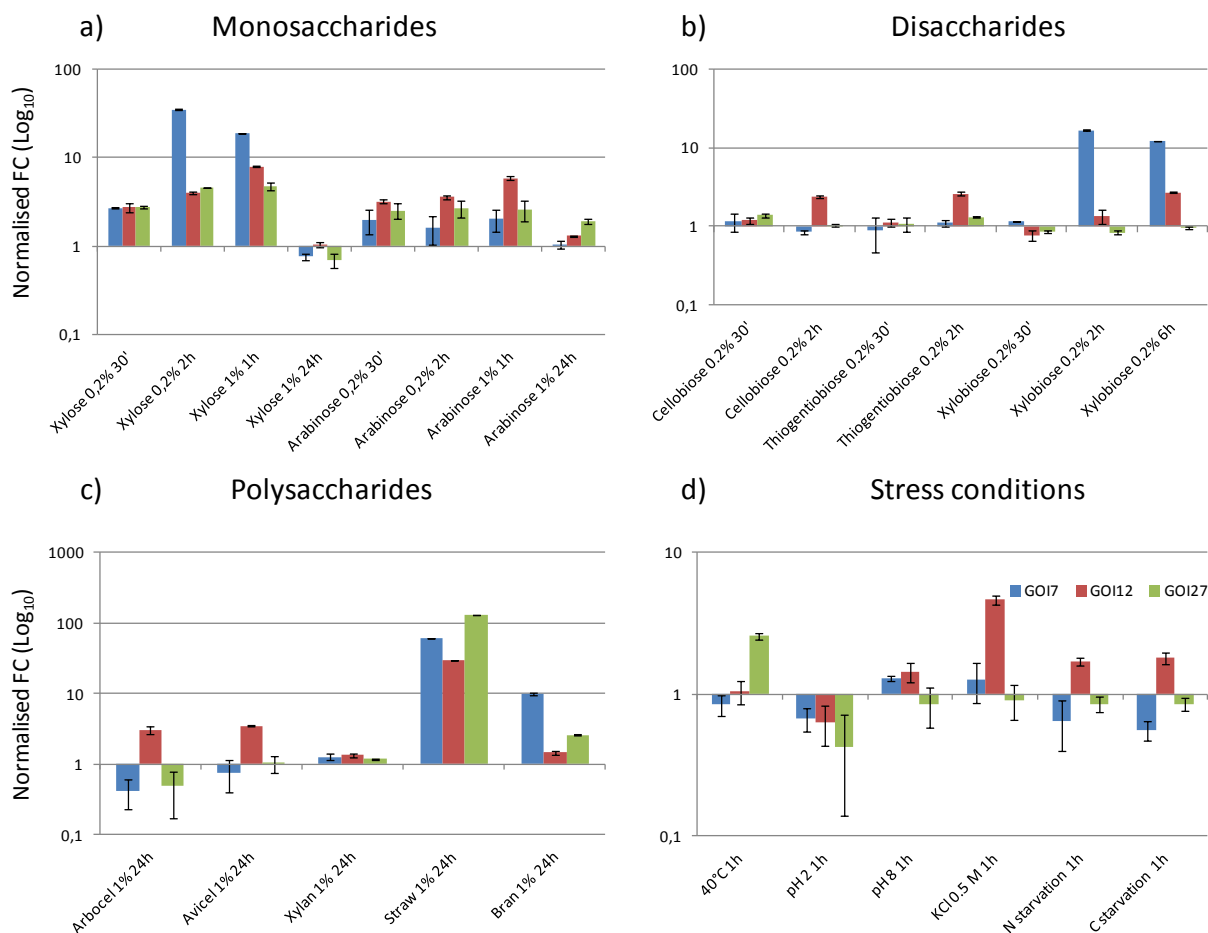


Figure 5.23: qPCR results for the class 3 genes. The *abf-B1* is represented by the blue bars, *xynC* by the red bars and the arabinofuranosidase GH43 by the green bars. The response of the genes was analysed when the fungus is exposed to monosaccharides (a), disaccharides (b), polysaccharides (c) and stress conditions (d) with different concentrations and time of exposure. The fold change (FC) is calculated considering the glucose condition as a reference. The expression values are represented in a logarithmic scale. Error bars represent SEM.

5.2.2.5. The class 4 expression pattern: the β -xylosidase genes

This class groups only the β -xylosidase genes, one of which belongs to the GH3 family (GOI19) and the other to the GH43 family (GOI23). These two genes, GOI19 and GOI23, were not well correlated with any of the other genes studied, as it was shown in the correlation matrix (section 5.2.2.1), but they show a good correspondence between the two of them.

The two monosaccharides tested, xylose and arabinose, enhanced the expression of both genes, although xylose in high or low concentration provoked a stronger response (up to 5,000-fold induction) than arabinose (up to 18-fold induction). In accordance with the observation made for the main hemicellulases (class 2), the induction grows in time when xylose is present in a small concentration (0.2%). The induction observed with arabinose is lower, but a low concentration of the substrate an increase in the induction over time is observed (Figure 5.24 a).

The two genes are induced by the disaccharides tested, but the pick of induction appears after 2 hours of exposure and not before. With cellobiose only 5-fold induction is observed and with thio-gentiobiose, 7-fold induction. Xylobiose is the best inducer in this group of substrates: it produces a response 50 times bigger than the cellobiose and the thio-gentiobiose that can reach 700-fold up-regulation (Figure 5.24 b).

Regarding the polysaccharides, both genes were significantly induced again. Wheat straw and xylan were the best inducers: they triggered an up-regulation between 250 and 800-fold. Even though the celluloses, Avicel and Arbocel, up-regulated the promoters of these two genes, the induction are significantly lower; between 13 and 67-fold (Figure 5.24 c).

The high temperature produced up to 7-fold repression of both genes. Acid pH only induced a 3-fold activation of the expression of one of the β -xylosidase genes (GOI23), which was also 2-fold induced with alkaline pH, but not the other gene (GOI19). The presence of salt stress or the nitrogen starvation were repressive conditions for this class. Nevertheless, the carbon starvation provoked a 3-fold induction of the GOI23 (Figure 5.24 d).

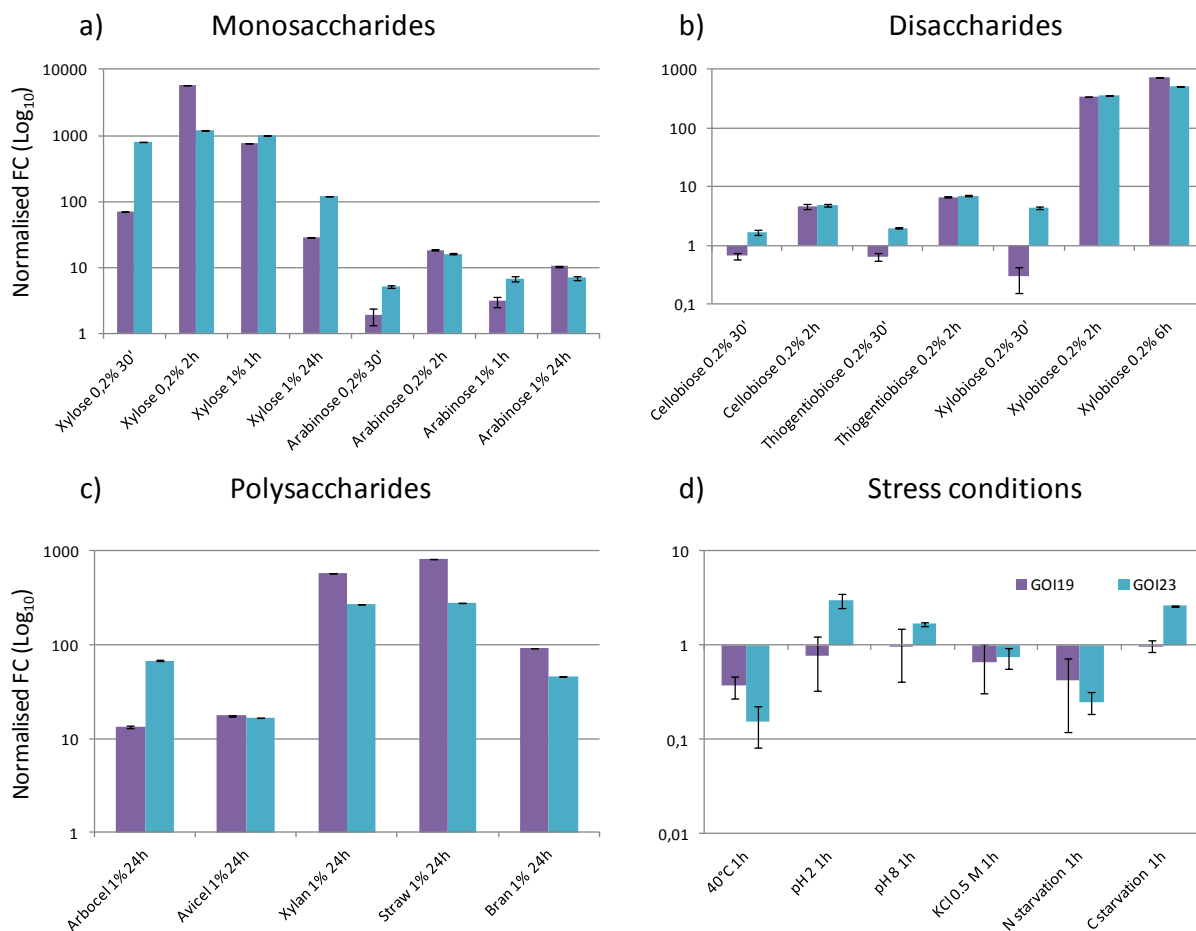


Figure 5.24: qPCR results for the class 4 genes. The β -xylosidase GH3 is represented by the violet bars and the β -xylosidase GH43 by the light blue bars. The response of the genes was analysed when the fungus is exposed to monosaccharides (a), disaccharides (b), polysaccharides (c) and stress conditions (d) with different concentrations and time of exposure. The fold change (FC) is calculated considering the glucose condition as a reference. The expression values are represented in a logarithmic scale. Error bars represent SEM.

5.2.2.6. *The class 5 expression pattern: the hydrophobin gene*

The last class is only composed by the hydrophobin encoding gene (GOI31). It has not been demonstrated that this protein is indeed involved in the degradation of the plant cell wall, but the expression of genes coding for different hydrophobins have been detected as differentially expressed in the RNA-seq study presented before (see section 5.1.1). Therefore, we decided to study in detail the expression of one of these genes to determine the regulatory mechanism that controls its

expression. When we performed the HAC analysis, the hydrophobin gene was classified with the hemicellulase genes, but the correlation matrix showed that the expression pattern of this gene was clearly different from all the other GOIs studied here, sharing very few characteristics with the other cellulase and hemicellulase genes.

The expression level of the hydrophobin encoding gene is down-regulated by the presence of xylose or arabinose in the medium, no matter what concentration (Figure 5.25 a). This gene can reach a 13-fold repression. Similarly, this gene is repressed by the different disaccharides tested, and the repression increases with time and it can reach a fold-change of 9 (Figure 5.25 b). In the case of the polysaccharides, the hydrophobin gene is strongly induced by the presence of wheat straw (8,000-fold) or wheat bran (12,000-fold) in the medium, but the cellulose or the xylan produce no change in the expression level (Figure 5.25 c). This suggests that the lignocellulosic material may contain a compound capable of inducing the expression of the hydrophobin through the action of a probably unknown transcription factor. A similar situation was previously observed for the gene coding the swollenin. Additionally, the carbon and nitrogen starvation strongly repress the expression of the hydrophobin encoding gene (Figure 5.25 d).

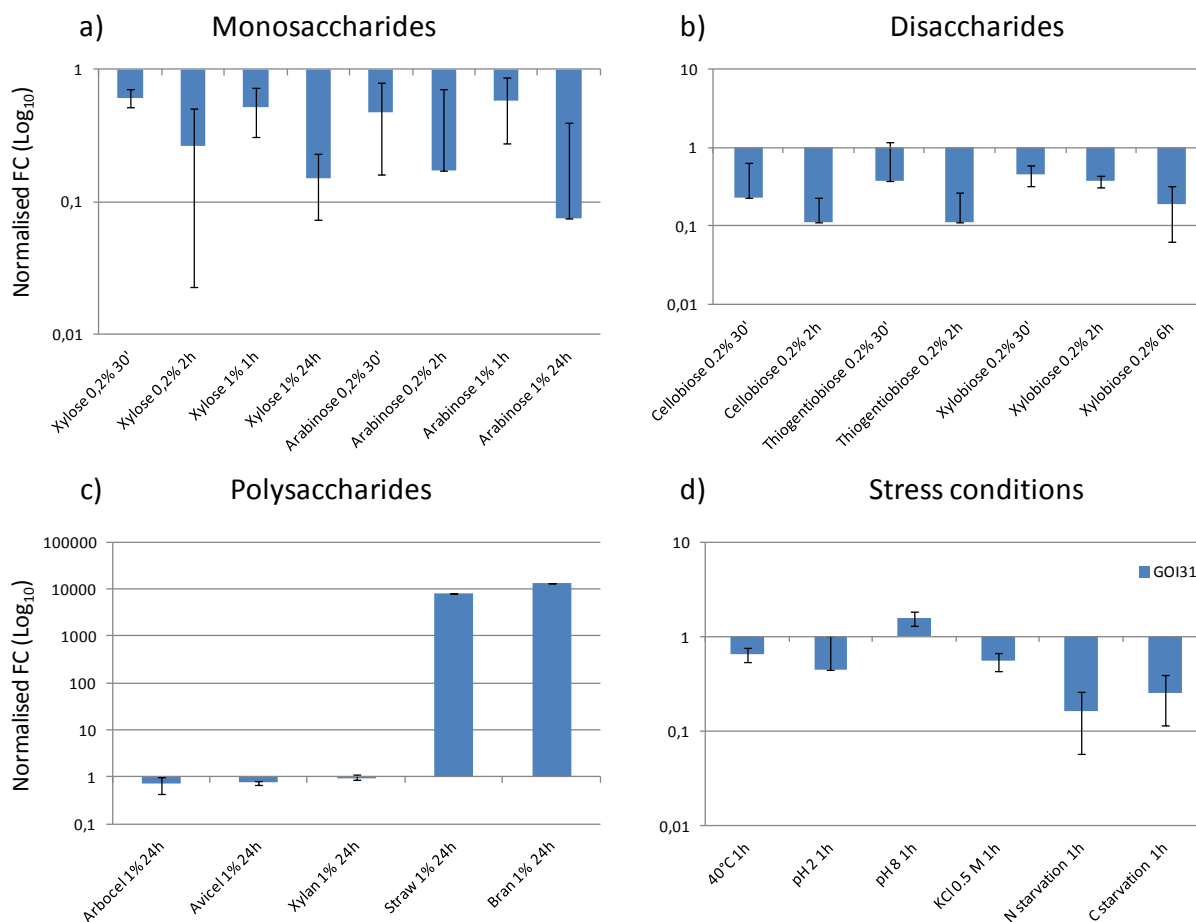


Figure 5.25: qPCR results for the hydrophobin gene. The response of the gene was analysed when the fungus is exposed to monosaccharides (a), disaccharides (b), polysaccharides (c) and stress conditions (d) with different concentrations and time of exposure. The fold change (FC) is calculated considering the glucose condition as a reference. The expression values are represented in a logarithmic scale. Error bars represent SEM.

5.3. Discussion

The genome-wide transcriptome study was carried over RNA samples obtained from three very different mycelia samples: mycelium exposed to glucose for 48h, then transferred to media containing ball milled wheat straw for 24h and finally the simultaneous exposure to wheat straw and glucose for 5h. This experiment allowed identifying the differentially expressed genes between the repressive conditions - in presence of glucose - and an inducing condition - in presence of wheat

straw. An enormous transcriptional change was observed between the conditions studied, in particular between Glucose 48h and Straw 24h. The two conditions containing glucose were not so different, which may indicate that the fungus prefers the glucose as an easily metabolisable compound, and that the use of other carbon sources might be rather inconvenient at an energetic level.

When the glucose is removed from the medium and is replaced by the complex lignocellulosic material wheat straw, the fungus needs to adapt to the new carbon source, that is to say that it needs to produce the machinery to degrade the polymers and incorporate the small molecules that are released from the degradation of the lignocellulose. As demonstrated by Zhao *et al.* (2013) the array of genes coding for hydrolytic enzymes varies depending on the fungus and its nutritional strategy, but in general they present a great diversity of hydrolase genes. The degradation of lignin, which represents 22% of the wheat straw, requires the presence of specialized enzymes such as peroxidases and laccases. This kind of enzymes is only distributed amongst the Basidiomycota (Sigoillot *et al.*, 2012). Hence, it is not surprising that the genome of *T. versatilis* does not present any gene coding for enzymes for the degradation of lignin (See chapter 4). *T. versatilis* is not prepared for the proper degradation of this aromatic compound. However, *T. versatilis* has a wide range of cellulases and hemicellulases encoding genes, and their expression is highly induced after 24 hours of incubation on wheat straw. The 48% of the inventoried GH-encoding genes were induced in the Straw 24h condition, many of the genes induced encode for similar proteins and in some cases several genes in the same CAZy family were induced. These enzymes may have different levels of efficiency and may act in different circumstances or locations in the same substrate. The necessity for so many similar enzymes may be justified by the synergistic action of all of them in order to efficiently degrade the plant cell wall polymers. In fact, more than 13% of the transcriptome of *T. versatilis* on wheat straw corresponds to hydrolytic enzyme genes mRNA. Which compared to the 0.15% of the transcriptome that the same genes represent in the Glucose 48h conditions, it is actually a vast change and an important investment for the degradation of the plant polymers. This is

in accordance with the findings of (Delmas *et al.*, 2012) and Ries *et al.* (2013) on *A. niger* and *T. reesei*, respectively. Even if the regulatory mechanisms are not necessarily the same between the three fungi, the transcriptome profile during the exposure to lignocellulose is similar.

Within the mRNA that corresponds to the GH-genes in each condition, it was observed that a great rearrangement was produced upon the exposure to wheat straw, where a deep switch in the proportions of mRNA that represent the different CAZy families was observed. The GH3, GH13 and GH31 encoding genes were the ones that have the highest percentage of mRNA in the Glucose 48h condition. But when the mycelium was switched to Straw 24h, the GH11, GH7, GH5, GH6, GH10, GH45 and GH54 took over (Figure 5.6). It is important to remember that these percentages are within the GH-genes mRNA only, and do not concern the other genes in the genome. This means that even if the genes from family GH31 are not predominant any more in Straw 24h, they are not necessarily repressed, but the amount of mRNA coming from these genes is small compared to the mRNA produced from the genes in family GH11, for instance. As a matter of fact, some of the genes of families GH31 and GH13 are stably expressed and some others are induced, but of course, this induction does not reach the levels of induction detected for the xylanase and arabinofuranosidase genes from other CAZy families. The CAZy families overrepresented in the transcriptome in the Straw 24h condition provide all the necessary enzymes for the full and efficient degradation of the lignocellulose, GH11 and GH7 being the ones that bear the highest mRNA proportions. Similar observations were done in the transcriptome of *A. niger* and *T. reesei* (Delmas *et al.*, 2012; Ries *et al.*, 2013).

The addition of glucose in the third condition, Straw + Glucose 5h, provokes another turnover in the transcriptome. Hence, the proportions of the mRNA corresponding to the different CAZy families changes one more time. After 5 hours of incubation in this condition, a more equal distribution of the mRNA of the different families is observed in this condition. This may indicate that a full repression of

the cellulase and hemicellulase genes could not be achieved in such a short time, or that the effect of the inducers could not be completely removed.

The genes coding for proteins such as swollenins and hydrophobic binding proteins were induced in the Straw 24h condition. The induction of these genes may be due to the fact that they are involved indirectly in the degradation of the lignocellulose, somehow improving the interaction of the other enzymes or the fungus itself with the substrate and assuring an efficient deconstruction of the polymers. As discussed previously, the hydrophobins from *A. oryzae* are involved in the degradation of the PBSA (Ohtaki *et al.*, 2006), and the ones from *Trichoderma* species, in the degradation of the PET (Espino-Rammer *et al.*, 2013). Even though the function of these proteins in *T. versatilis* has not yet been elucidated, a similar model could be considered. Moreover, we detected the induction of an important number of genes coding for transporters, which may be involved in the transport of sugars into the cell, and genes encoding enzymes with esterase activity, that may also be involved in the degradation of complex polymers.

Amongst the known transcription factors controlling the expression of GH-encoding genes, only *xlnR* and *clrA* were found to be significantly differently expressed. The induction of *xlnR* may be the result of the release of CreA in the *xlnR* promoter, as it was described in *A. nidulans* (Tamayo *et al.*, 2008) and *T. reesei* (Mach-Aigner *et al.*, 2008). This assumption is not in accordance with the findings made by Mach-Aigner *et al.* (2012) in *A. niger*, in which the transcript levels on the *xlnR* gene are constant and independent of the presence of CreA. However, this data supports the fact that the regulatory mechanisms may differ in the different fungi. The transcription factor *hacA*, responsible for the unfolded protein response, was also induced in Straw 24h, but it was not considered to be significantly differently expressed. However, the increase in the expression of this transcription factor is consistent with the fact that in this situation there is an important amount of proteins being produced and entering the secretion pathway and probably a considerable amount of those proteins are misfolded and do not pass the quality control of the ER. In this condition the activation of the

UPR becomes necessary in order to cope with the stress provoked by the accumulation of unfolded proteins. More interestingly, within the differentially expressed genes there are the genes encoding for the BlrA and the AlcR transcription factors, that control the conidiation and the alcohol metabolism related genes respectively, and 11 other genes encoding for transcription factors with unknown function. The 11 unknown transcription factors could become interesting target for further study, since they might be directly involved in the induction of the GH-encoding genes and could contribute to the extremely high up-regulation of these genes in the Straw 24h condition.

The new annotation of the genome contributed to the improvement of the quality of the mapping of the RNA-seq results. The new results were analysed using a more performant pipeline in order to find the significantly differently expressed genes: DESeq. A much bigger number of differentially expressed genes were determined with this method, compared to the DEGseq used previously. The GO terms enriched within the differentially expressed genes gives us an idea of the state and the changes in the cells between the conditions under study. It was predictable to see the polysaccharides catabolism, cellulase, hydrolysis and xylanase activities, and even the pectin catabolism and the transport GO terms enriched when the fungus is exposed to wheat straw for 24 hours. These are the functions that the fungus needs to degrade the complex substrate to which it is confronted. But it was surprising to see that in Straw 24h the down-regulated GO terms were those related with the ribosome synthesis, rRNA processing and maturation, amino-acids-synthesis, mitochondrial functions, ATP synthesis, proton transport and mycelium development. This leads us to think that in Straw 24h the fungus is in a “steady state”, where all the main cellular functions are still there, but no development is carried on. This might be a way to cope with the stress caused by the lack of a favourable carbon source. This entire situation is reverted from the addition of glucose to the medium again.

Quantitative PCR was used to study in depth the expression pattern of a group of genes of interest that included cellulases, hemicellulases and other auxiliary proteins encoding genes. We

wanted to investigate the expression of these genes when the fungus is exposed to different complex substrates, but also the monosaccharides and disaccharides that are produced when the polymers are degraded in nature and that are supposed to be the real inducers of the expression of the hydrolytic genes. A series of stress conditions were also tested, including pH and nitrogen or carbon starvation. In a first stage, the expression of all the genes in all the conditions was analysed in general through a hierarchical classification of the genes that allowed the division of the 15 genes in study in 3 distinct classes. Each of these classes contains different kinds of enzymes whose expression patterns are similar. Also, we observed that their transcript level depended on the kind of condition to which the fungus was subjected. It is remarkable how genes coding for similar enzymes have very different expression patterns and are split apart in the classes, like the *xynC* and *xynB*, which belong to the same GH family, but are expressed in completely different ways. It is possible that they accomplish different functions and are necessary in different circumstances; therefore their promoters are activated or repressed by different transcription factors.

However, the hierarchical classification does not take into account the correlation of the expression in the different conditions, that is to say, if two different genes are up- or down-regulated in response to the same substrates. Thus, the correlations between different genes were studied and the genes with similar behaviour in all conditions were grouped together considering only the highest correlation coefficients. In this way, five new classes were defined, and within the classes, the genes show very similar expression patterns and the groups have a biological significance. The five classes were composed of different kinds of genes: the class 1 included the cellulase genes; class 2 consists of the hemicellulase genes that were highly up-regulated in most conditions; class 3 puts together the other hemicellulase genes that had rather low expression levels; class 4 only includes the two β -xylosidase genes; and class 5 is the hydrophobin gene. It would be worth considering that the promoter regions of the genes in the same class may share regulatory elements and that they might be controlled in a similar way. The comparison of the promoter regions could be considered as a way

to get a better understanding of the regulatory networks that play a role in the expression of these genes.

All the genes studied with qPCR were strongly induced by the presence of wheat straw in the medium, which is consistent with the transcriptomic results. However, when comparing the results from qPCR with the RNA-seq data in detail, the levels of induction did not always correlate. For instance, on qPCR the *xynD* gene is much more induced than the *xynB* (215,459-fold and 3766-fold, respectively), whereas in the RNA-seq data, the transcript level values suggest that *xynD* is less induced than *xynB* (459-fold and 407-fold, respectively). A similar situation is observed for the cellobiohydrolase genes: *chbII* induction is higher than *chbI* in qPCR (10,814-fold and 4,238-fold), but the RNA-seq data suggest that they are induced at the same level (426-fold and 418-fold). The different techniques used for the estimation of the amount of mRNA in a sample could be the origin of these differences, also the preparation of the libraries could have influenced the final result. More importantly, the number of replicates has an important influence in the accuracy of the results. Since the RT-qPCR results are preliminary results, with only 2 biological replicates, a bigger number of replicates should be performed to reduce the stochastic effect. Moreover, the RNA-seq results used for this analysis correspond to the one obtained from the alignment to the old genome annotation. The alignment of the reads to this genome was not perfect and this may have led to errors in the calculation of the RPKM values. This could explain the discrepancies in the data. An analysis of the same data aligned to the new genome annotation may present different results, further work on the RNA-seq data now available is due to be done.

Nevertheless, the induction of all the GOI when the fungus is exposed to wheat straw was confirmed through the quantitative PCR experiments. In fact, the wheat straw is the substrate that leads to the highest induction of them all. It seems that the highest the complexity of the substrate, the highest the induction of the GH-encoding genes, which could be the result of a synergistic action of several inducers together.

When looking at the expression levels of the genes of interest in presence of xylose or arabinose, we notice that a low concentration can induce the expression of most of the genes studied and that xylose is more efficient than arabinose for the induction of this group of genes. But when using a high concentration of these compounds an induction is observed after an hour of incubation, but a repressive effect can be observed in the long term. This is in accordance with the Mach-Aigner *et al.* (2012) results, where they demonstrated that in *A. niger* a high concentration of xylose induces the expression of *xynD* but only for the first 2 hours. They concluded that CreA is responsible for the repression of the genes in presence of high concentrations of xylose. Another example of activation first and repression later by high concentrations of sugar was described by Perez-Gonzalez *et al.* (1998) in *A. nidulans*. The carbon catabolite repression could respond not only to glucose, but to other simple sugars, if they are in a sufficiently high concentration. In this kind of situation, the fungus does not need to degrade polymers, since it is easier to uptake and metabolise the monosaccharide, therefore the expression of the cellulase and hemicellulase genes are down-regulated.

Since the polysaccharides cannot enter the cell, it is the products of its degradation that enter and act as inducers for the expression of the hydrolytic enzymes, so that the full degradation of the polymers can be achieved. For instance, cellobiose was described to induce the gene expression in *A. nidulans* (Chikamatsu *et al.*, 1999) and *N. crassa* (Znameroski *et al.*, 2012). From our data, we observed that the genes are positively regulated by the presence of cellobiose or xylobiose. In fact, the hemicellulase genes respond better to xylobiose than to cellobiose, whereas the cellulase genes show a stronger induction when cellobiose is present in the medium, than in presence of xylobiose. This kind of compounds is easily degraded by the β -glucosidases, in the case of the cellobiose, and β -xylosidases, in the case of the xylobiose, and then metabolised. Thus, an induction sustained in time is not easily achieved. This is probably the reason why the transcript level of the cellulase genes decreases after 6 hours of incubation in xylobiose. The induction of the hemicellulase genes is

maintained after 6 hours of incubation in xylobiose; an induction produced by the xylose released from the degradation of the xylobiose could be the explanation to this effect.

All of the polymers tested induced most of the genes studied, except for those that belong to the classes 3 and 5, which are only induced by the presence of complex lignocellulosic material, wheat straw or wheat bran. The main hemicellulase genes from class 2 responded to both, the cellulose and the xylan. The cellulase genes (class 1) as well, but the induction by xylan was lower. In *N. crassa*, the Clr1 transcription factor is involved in sensing the presence of cellobiose, the degradation product of the cellulose and it can activate the necessary genes for the import and metabolism of cellobiose. It also activates indirectly, through Clr2, the expression of cellulase and hemicellulase genes (Coradetti *et al.*, 2012). Probably, the induction of the hemicellulase genes is mediated by XlnR and ClrA at the same time, whereas the cellulase genes are mostly controlled by ClrA. That could also explain the extremely high levels of induction of the hemicellulase genes in presence of any kind of cellulose. The wheat straw and wheat bran proved to be the best inducers, more likely because of the mixture of substrates that they contain; all the inducers can act in a synergistic manner and reach the full activation of the GH-encoding genes. Interestingly, the hydrophobin gene was only activated by these two substrates, suggesting that this gene responds to an unknown inducer present in the wheat straw and in the wheat bran. The difference between the pure cellulose, the xylan and the wheat straw and wheat bran is the presence of lignin and pectin. The inducers of this gene may be within these substrates. No genes coding for enzymes for the degradation of lignin are present in the *T. versatilis* genome, but several pectinolytic genes have been inventoried, like the genes that belong to the families GH43, GH54, GH51. Further studies are necessary to determine the regulatory mechanism that governs the expression of the hydrophobin gene.

The nitrogen starvation did not induced any of the genes tested, suggesting that the homologous of the AreA regulator present in *T. versatilis* does not have the same function than the one described in *A. nidulans*, as activator of the cellulase encoding genes (Lockington *et al.*, 2002). The carbon

starvation, on the other hand, induced some of the genes studied and the induction observed is relatively low. This is consistent with the model described in *A. niger* (Delmas *et al.*, 2012) and in *N. crassa* (Tian *et al.*, 2009), where if the fungus is exposed to a complex carbon source that cannot be incorporated, it senses it as a lack of carbon and will produce hydrolytic enzymes at a low level by de-repressing the promoter regions of these genes. The enzymes start the degradation of the substrate and the molecules released will induce the full expression of the hydrolytic enzymes genes. The moderate induction of some hydrolytic enzymes in a carbon starvation condition is part of this adaptive response. The alkaline pH also slightly induced the expression of some of the genes studied, including the cellobiohydrolase genes, the *abf-B2*, the *xynD* and the *xynC*, suggesting that they might be under the control of the PacC transcription factor. The acid pH and the high temperature were repressive conditions for the genes of interest.

6. Effect of the deletion of key transcription factors over the expression of the GH encoding genes

6.1. Construction of the deletion mutants

The tools for genetic modification of *T. versatilis* were very limited. A protoplasts mediated transformation had been adapted to this fungus prior to the start of this thesis, but the DNA modification itself needed to be improved. A method for gene deletion in filamentous fungi based on recombination between a plasmid and the chromosome was designed for *A. niger* and then adapted to *T. versatilis* during this thesis (Delmas *et al.*, 2014).

6.1.1. Construction of a *Talaromyces versatilis* Δ pyrG strain

To apply this method, it was first necessary to construct a host strain with a deletion in the *pyrG* gene. The *pyrG* genes encodes the orotidine-5'-monophosphate decarboxylase protein, an enzyme involved in the nucleotide biosynthesis that catalyses the synthesis of the orotidine-5'-monophosphate (Wittmann *et al.*, 2008). The lack of this enzyme prevents the growth in media without uridine. The deletion of this locus was used in different fungal strains as the genetic background for other DNA modifications (Fierro *et al.*, 2004; Navarrete *et al.*, 2009; Nielsen *et al.*, 2008, 2006). In order to obtain the *T. versatilis pyrG* deletion mutant, 1614 bp of the upstream region and 1880 bp of the downstream region of the *pyrG* locus were PCR amplified and ligated. The linear fragment was used to transform the *T. versatilis* protoplasts (for more details on the protocol, see chapter 2). The cassette was integrated into the genome at the *pyrG* locus by double homologous recombination, leading to the elimination of the *pyrG* open reading frame on the chromosome, as it is shown in Figure 6.1.

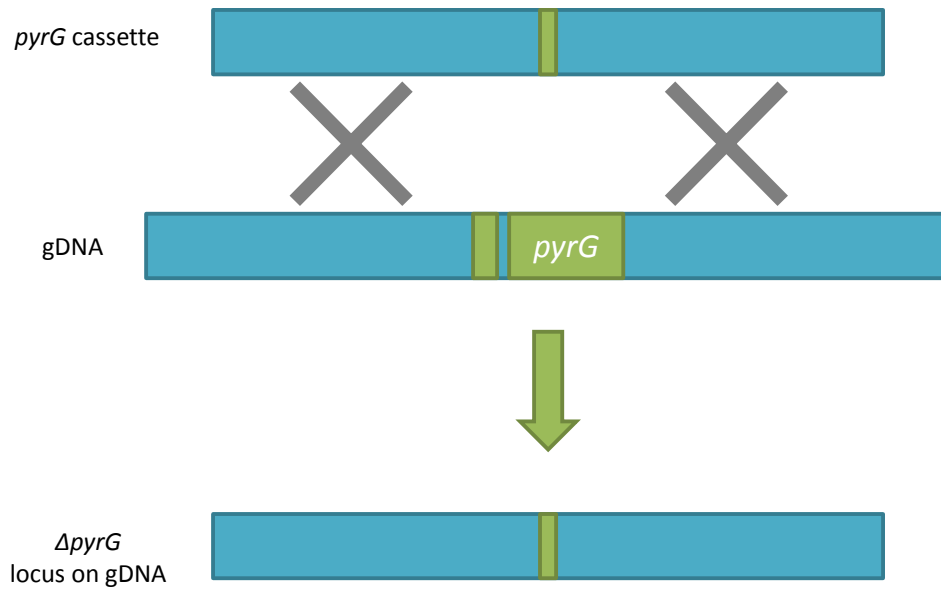


Figure 6.1: Schematic representation of the construction of *pyrG* deletion mutant strain of *T. versatilis*. The *T. versatilis* protoplasts were transformed with the cassette containing the upstream and the downstream fragments of the *pyrG* locus. The linear DNA fragment is integrated into the genome by a double homologous recombination event (grey crosses) creating a uracil auxotrophic strain.

For the transformation, 1×10^7 protoplasts were put in contact with $10 \mu\text{g}$ of the DNA cassette. After the transformation, the mutants were selected in media containing uridine 10 mM and 5'fluoroorotic acid (5'FOA) 1.25 mg/mL. The 5'FOA can be converted by PyrG to 5'fluorouracil, which is a toxic compound for the cell. The strains that lack the *pyrG* gene cannot grow on media without uridine and are insensitive to the presence of the 5'FOA. Only two transformants grew on this media. The deletion of the *pyrG* locus in both transformants was confirmed by PCR and sequencing of the locus, but only one of them was selected for further work. Thereafter, to probe that the strain is auxotrophic, the growth on plates with and without uridine was tested and, as it is shown in Figure 6.2, in absence of uridine, the growth is very limited, whereas in presence of 10 mM of uridine the strain presents a normal growth, even in the presence of 5'FOA.

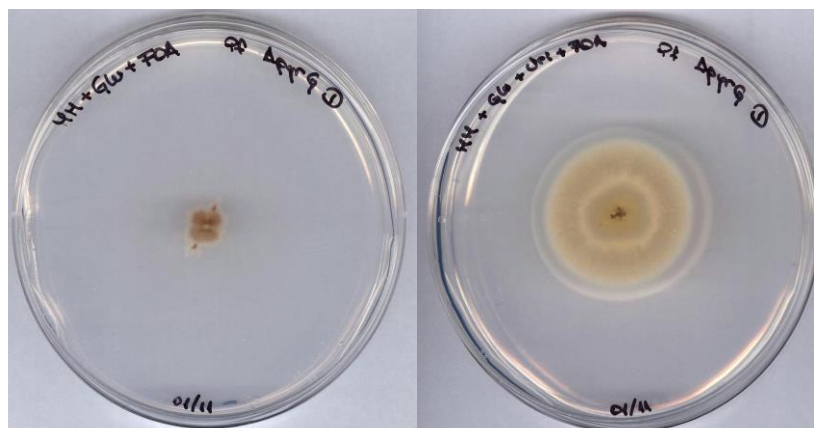


Figure 6.2: TvMM + glucose 1% + 5'FOA 1.25 mg/mL agar plates with growth of the *T. versatilis* $\Delta pyrG$ fungal strain inoculated at the centre. The left plate has no uridine, and the right plate was supplemented with 10 mM uridine. The plates were incubated at 30°C for 7 days.

The effect of the lack of the *pyrG* gene was reverted by the introduction of a heterologous *pyrG* gene. A pBluescript SK⁻ plasmid (Stratagene) containing the *pyrG* gene from *A. oryzae* with its promoter and terminator regions was used. The upstream and downstream regions of the *T. versatilis pyrG* locus were cloned in the multiple-cloning site of this plasmid to allow the integration into the genome by homologous recombination at the *pyrG* locus. The plasmid is represented in Figure 6.3. The *T. versatilis* $\Delta pyrG$ strain did not present conidiation in the LYMM plates usually used for the production of conidia. Therefore, conidia were produced in PDA plates supplemented with uridine 10 mM and thereafter protoplasts were generated following the usual protocol (see chapter 2). The transformation was done in duplicate, each was performed with 1×10^7 protoplasts and one of them with 10 μ g of plasmid DNA and the second with 6 μ g. The transformants were selected in TvMM + glucose 1%, but without uridine or 5'FOA. The two transformations originated 21 and 10 transformants, respectively. There were no colonies in the negative control transformation plates, which was done without any DNA. The transformation efficiency was 2.1 and 1.7 transformants per μ g of DNA for each experiment respectively, which is higher than the transformation efficiency obtained with the linear DNA (0.2 transformants/ μ g of DNA). This experiment allowed probing that

the plasmid could be taken up by the cell with a better efficiency than a linear DNA fragment and that the *pyrG* gene from *A. oryzae* is functional in *T. versatilis*, therefore it can be used as a selection marker for future genetic modifications.

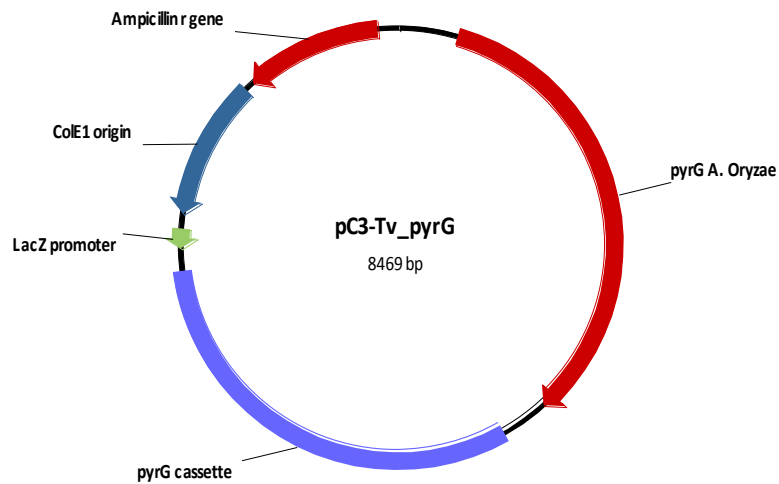


Figure 6.3: Plasmid pC3-Tv_pyrG. The pyrG cassette (light blue box) contains the upstream and downstream regions of the *T. versatilis pyrG* locus and it was cloned in the multiple-cloning site of the plasmid. Component parts of the vectors are indicated. The red box represents the *A. oryzae pyrG* locus that includes the ORF, the promoter region and the terminator region.

6.1.2. *Development of an unmarked gene deletion system for Talaromyces versatilis*

A method for gene deletion based on the homologous recombination between a plasmid and the genome was adapted for *T. versatilis* (Delmas *et al.*, 2014). The *pyrG* gene was used as a selectable marker. The advantage of using this gene as marker is that it can be selected and counter-selected; in fact, cells lacking the PyrG enzyme cannot grow without exogenous uridine/uracil, while these cells are resistant to the toxicity of 5'FOA.

The *A. oryzae pyrG* locus, containing the ORF, the promoter and the terminator regions, was cloned in the pBluescript SK⁻ plasmid (Stratagene), in both orientations, generating plasmids pC3 and pC7. Only the pC3 plasmid was used for the deletions in the *T. versatilis* industrial strain, which from

now on will be called “wild-type” strain. The plasmids retain the blue/white screening capability for cloning and 4 unique restriction sites (EcoRI, NotI, SpeI, and XhoI) in the multiple-cloning site to clone the flanking sequences of the locus to be deleted. The plasmid is not capable of replicating in filamentous fungi, since it has no origin of replication.

The plasmids for the deletion of 2 different transcription factors in *T. versatilis* were constructed: *xlnR* and *clrA*. The plasmids are shown in Figure 6.4. For the deletion of *xlnR*, 1791 bp upstream and 1613 bp downstream fragments were cloned into the pC3, and for the deletion of *clrA*, 1484 bp upstream and 1562 bp downstream fragments were cloned in the same plasmid.

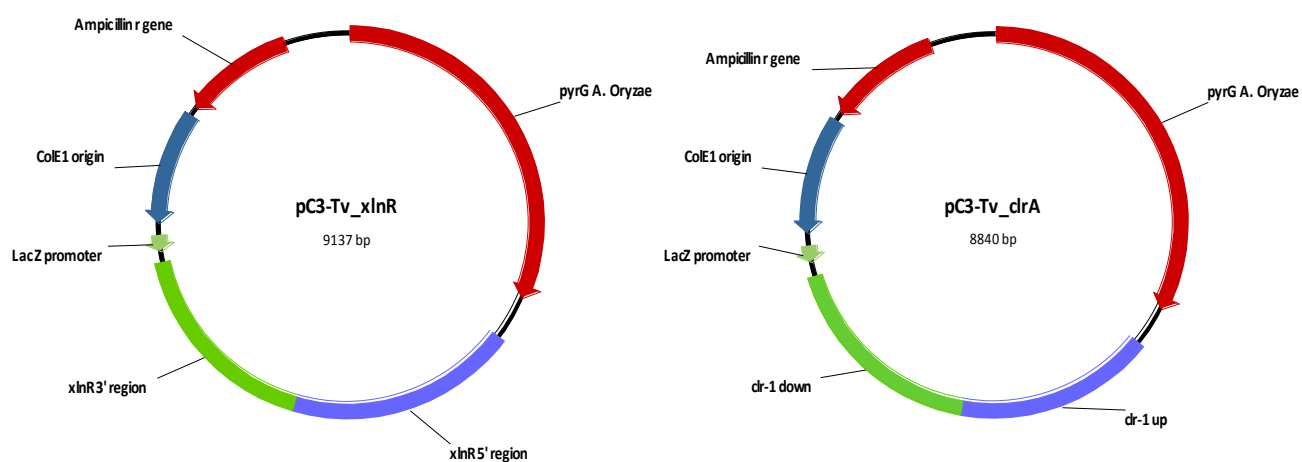


Figure 6.4: Representation of the pC3-*Tv_ΔxlnR* plasmid (left panel) and the pC3-*Tv_ΔclrA* (right panel). Flanking regions of *xlnR* are shaded: upstream in dark gray and downstream in light gray.

Transformations with these two plasmids were carried out. Once the DNA was taken up by the cells, the integration of the plasmid into the genome can occur by HR with the upstream or the downstream region. The recombination and integration with the upstream region is represented in Figure 6.5; now a functional *pyrG* gene is present in the genome. The transformants can be selected in media without uridine. Thereafter, the selective pressure can be released and a new recombination event will excise the plasmid from the genome. Depending on where this

recombination occurs, between the upstream fragments (blue) or the downstream fragments (green), it will generate a deletion mutant in the target gene or a wild type strain. After the excision of the plasmid, the strain no longer has the *pyrG* gene and the uridine auxotrophy is recovered. In the case of the deletion of *xlnR*, two transformations were performed; each gave 20 and 25 transformants, respectively. After the release of the selective pressure, 9 colonies were chosen and screened for the deletion of *xlnR* by PCR using primers external to the gene. Five of them showed the 673 bp band corresponding to the deleted locus, the other 4 had the 3470 bp band that is consistent with the wild type locus. The deletion mutants were also confirmed by sequencing of the *xlnR* locus. For the deletion of *clrA*, 3 transformations were done: the first one gave 28 transformants, the second one 30 and the third one 24. After the release of the selective pressure, a large screening was done in order to find a deletion mutant: in total 57 transformants were screened by PCR, only 9 presented the deleted locus and all the others the wild type locus. Two of the mutants were confirmed by sequencing of the *clrA* locus.

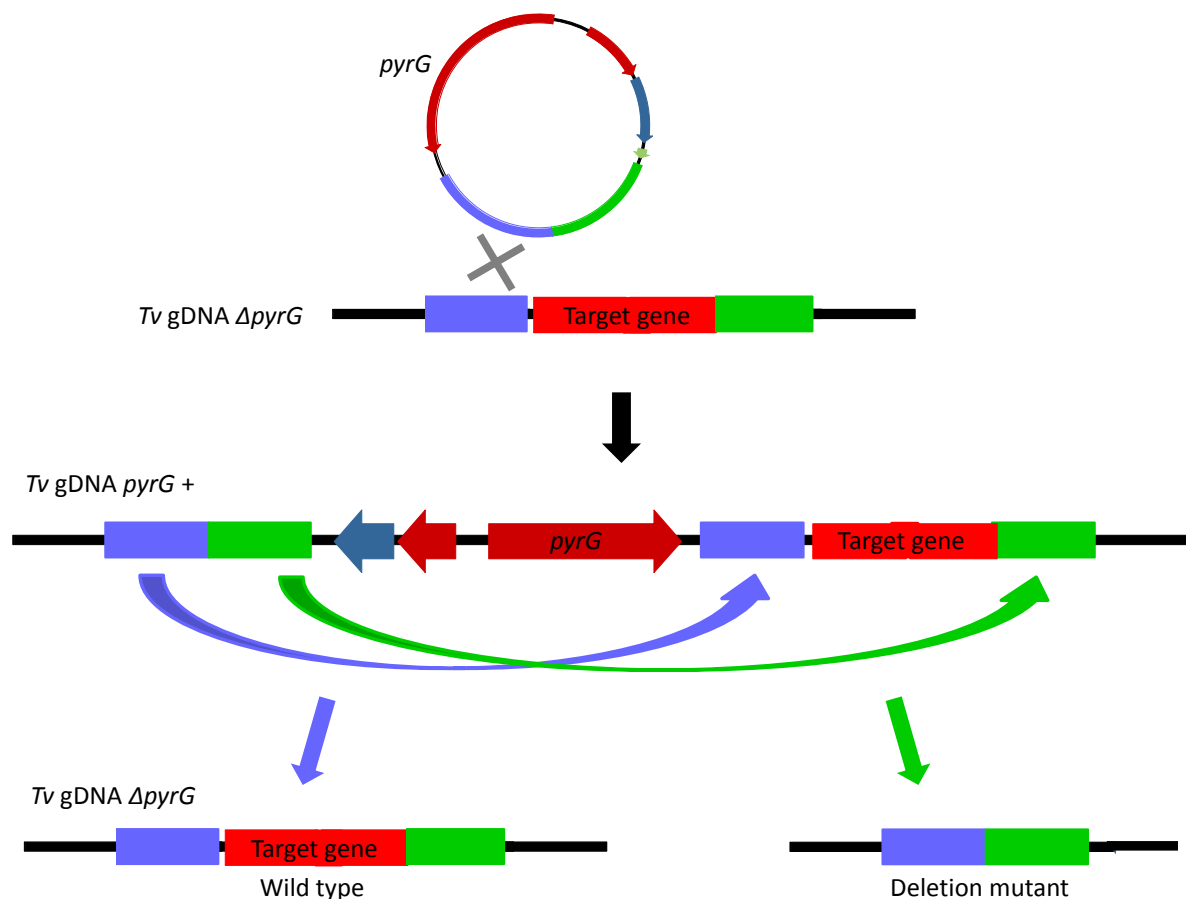


Figure 6.5: Scheme for deletion of a target gene by the intermediate of pC3 plasmid. First the recombination between the plasmid and the chromosome using the homology of the flanking region of the target gene present on the chromosome and the plasmid lead to the integration of the plasmid on the chromosome. Transformants that have integrated the plasmid are *pyrG*⁺ and can grow on medium lacking uridine. Releasing of the selective pressure for plasmid integration by growing cells on a medium containing uridine and 5'FOA allows the excision of the plasmid. When the plasmid is excised and cured, the cells are $\Delta pyrG$ and can be selected by plating on a medium containing uridine and 5'FOA. Excision using the homology of the flanking region can lead either to reversion to the wild type copy of the gene (on the left) or to the deletion of the locus (on the right).

The deletion of two other transcription factors were also tried, but without success: *araR* and *creA*. In the case of *araR*, the transformation was done twice, but we did not succeed to find a deletion mutant during the screening. For *creA*, on the other hand, two different transformations were tried. Firstly, we tried to delete the full length gene, but the transformants obtained were not

able to grow after the passage to fresh plates. Since it has been reported that the deletion of *creA* may be troublesome, we decided to delete the region coding for the zinc finger domain only. In this case, one transformation was done and it gave 6 transformants. Unfortunately, the screening for the deletion mutant was not finished at the end of the thesis works. The strains with the plasmid integrated into the genome are available to continue the screening for a deletion mutant.

6.2. Deletion mutants characterisation

The strains obtained with the previous method were affected in transcriptional factors that are intimately implicated in the expression of genes coding for glycoside hydrolases.

The deletion mutants obtained show particular phenotypic characteristics. For instance, both the *T. versatilis* $\Delta xlnR$ and the *T. versatilis* $\Delta clrA$ strains grow slowly in liquid TvMM. When *T. versatilis* $\Delta xlnR$ grows on PDA plates, the mycelium also produces the deep red stain observed in the wild-type strain and the conidia that are produced have the same brownish colour than the wild-type strain. However, we observed that the conidia from the *T. versatilis* $\Delta xlnR$ have a much smaller size than the wild-type *T. versatilis* strain. Conversely, the *T. versatilis* $\Delta clrA$ strain produced a very thick and compact mycelium with a granulated surface when cultivated in PDA. Like the wild-type strain, it also produced the deep red pigment that stains the agar. The conidia take a long time to appear on this media, they are visible only after 15 days of incubation and they are scarce. It is possible that this strain can sporulate better in other media than PDA. The appearances of the 3 strains after 10 days of incubation in PDA slopes at 30°C are shown in Figure 6.6.

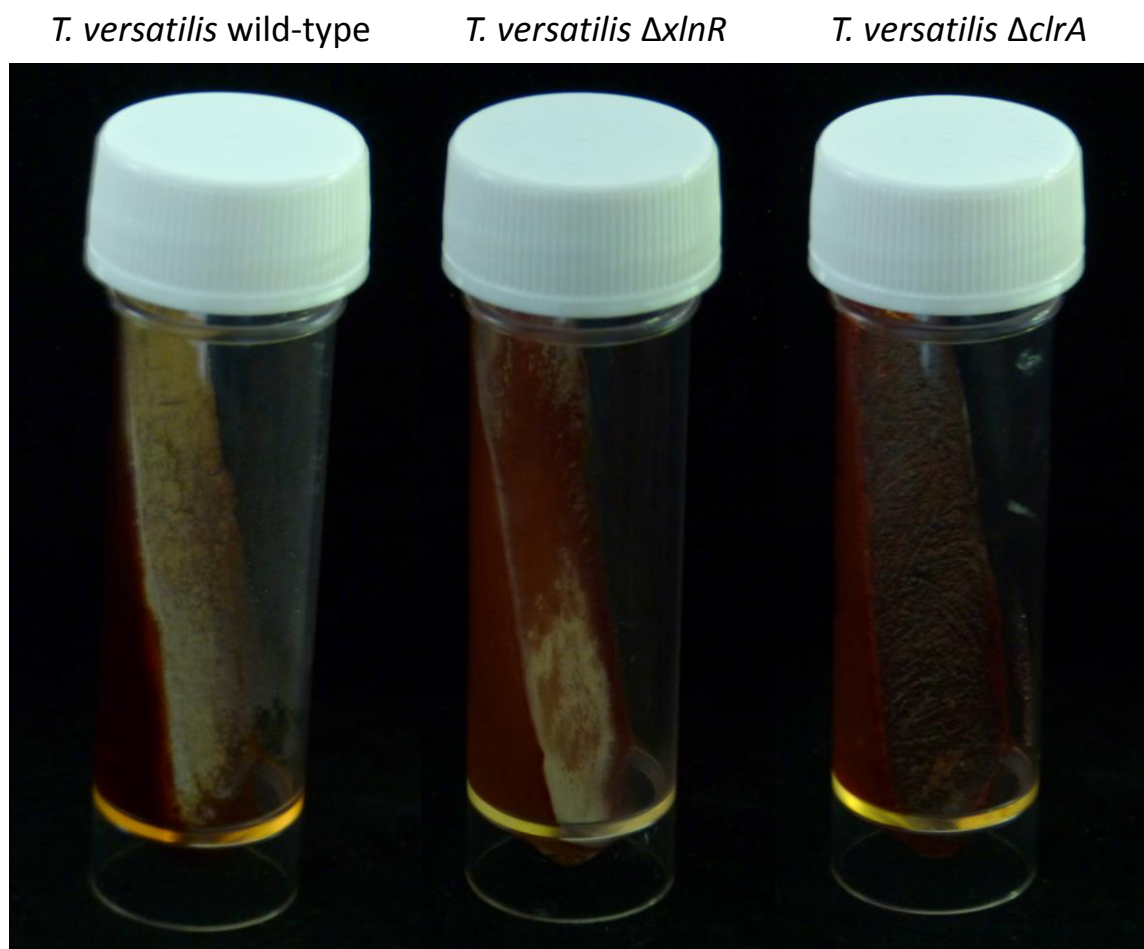


Figure 6.6: Growth of *T. versatilis* strains on PDA slopes. The wild-type strain is shown on the left, the $\Delta xlnR$ strain in the centre and the $\Delta clrA$ strain on the right. Pictures were taken after 10 days of incubation at 30°C.

In order to determine qualitatively how the deletion of such transcription factors affected the production and secretion of GH proteins, a congo red test was carried out (see detailed protocol in Chapter 2). For this test, the wild-type *T. versatilis* strain and the two deletion mutants $\Delta xlnR$ and $\Delta clrA$ were cultivated in TvMM supplemented with 1% glucose as carbon source, the mycelia produced was washed and transferred to TvMM with wheat straw 1% to induce the secretion of the enzymatic cocktail. The three strains were incubated for 24h and the supernatant containing the enzymes was recovered. Drops of 10 μ L of supernatant were deposited in the centre of a plate containing CMC agar or Xylan agar. After 2h of incubation, the plates were stained with congo red.

The congo red is a water soluble dye that has affinity for the cellulose fibrils as well as other polysaccharides, but it cannot bind the monosaccharides. Hence, the halo visible at the place where supernatant drops were spotted is due to the degradation and disappearance of the polymers by the action of hydrolytic enzymes (see Figure 6.7, centre of the drop indicated by a black cross).

The supernatant of the wild-type strain is capable of degrading both the carboxymethyl-cellulose (CMC) and the xylan, the halos are approximately 13 mm and 12 mm, respectively. The *T. versatilis* $\Delta clrA$ strain, on the other hand, shows no degradation whatsoever in any of the substrates. Only a very faint halo was visible to the naked eye and it was located in the spot where the drop of supernatant was put. And finally, the supernatant from the *T. versatilis* $\Delta xlnR$ strain presented a slightly reduced halo in CMC, 11 mm diameter. But the halo in xylan was notably reduced to 8 mm. All these results taken together suggest that the XlnR regulator is involved in the expression of the xylanase encoding genes and, to a lesser extent, in the expression of the cellulase genes. But the ClrA regulator has a major function in the expression of both the cellulases and xylanases. Further experiments are necessary to determine to what extent these regulators are essential to the expression of the GH-encoding genes.

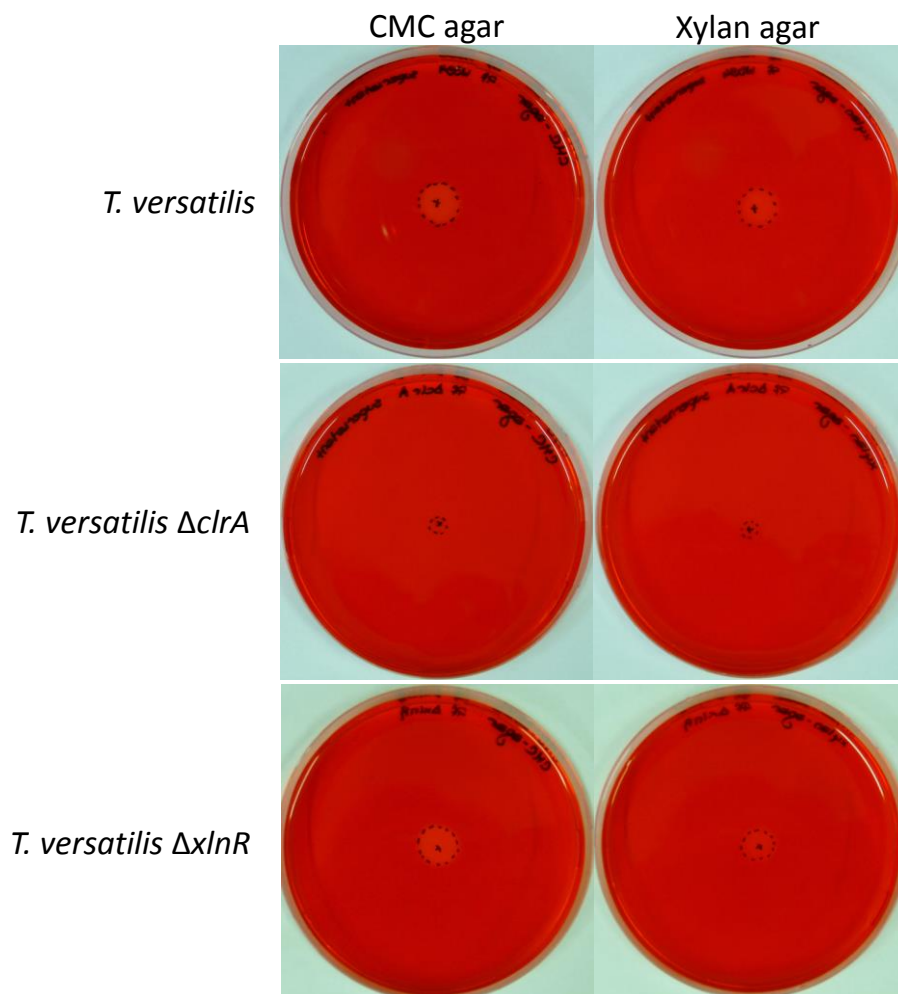


Figure 6.7: Halos of hydrolysis in CMC agar (left) and Xylan agar (right) plates produced by the wild-type *T. versatilis*, *T. versatilis* $\Delta clrA$ and *T. versatilis* $\Delta xlnR$. The cross in the middle of the plate indicates the spot where the supernatant was deposited. The dotted line around indicates the limit of the halo. Notice that for the *T. versatilis* $\Delta clrA$ strain, the halo was faintly visible to the naked eye and probably no degradation occurred.

6.3. *Study of the expression levels of the of genes of interest in a selection of substrates*

The expression of the 15 GOIs studied in the *T. versatilis* wild-type strain (see Section 5.2) were also analysed in the *T. versatilis* $\Delta xlnR$ strain. The transcript level of these genes was measured in 10 different conditions, including the exposure to the polymers Avicel, xylan and wheat straw, the

carbon and nitrogen starvation and the exposure to arabinose and xylose in low (0.2%) and high (1%) concentrations. All the normalised expression values obtained are shown in the Appendix VIII.

The distribution of the transcript levels of each gene in the 10 conditions tested and in both strains, the wild-type (green boxes) and the *xlnR* mutant (violet boxes), are shown in the boxplot in Figure 6.8. Within the class 1, the mutant presents wider boxes than the wild-type strain, indicating a higher variability in the transcript levels. More interestingly, the transcription of GOI2, GOI4 and GOI33 reach a higher level in the mutant than in the wild-type strain. The GOI2 and GOI4 have fairly the same transcript level in both strains but the boxes are wider in the mutant. Even if the boxes have different sizes in all these genes, the median (indicated by a line in the box) and average (indicated by a cross) do not change dramatically. In class 2, the genes GOI6, GOI10 and GOI14 show lower transcript levels in the mutant than in the wild-type. In this class, the median and average values are lower in the mutant strain than in the wild-type, although the values are not statistically different. Within class 3, GOI7 and GOI27 show similar overall distribution in both strains, with similar median values as well. But GOI12 shows a much reduced distribution in the *xlnR* mutant, with very low transcript levels in general, as highlighted by the significantly lower median value. Therefore, with the exception of its low FC values, which is its specificity, GOI12 somehow resembles class 2 GOIs. The 2 genes in class 4 show a reduced variability in the mutant strain, indicated by the smaller boxes, but the median values remain similar in both strains. And finally, the hydrophobin gene (GOI31) in class 5 presents a higher variability in the mutant strain and a gain in the transcript levels when we look at all the conditions together. Even though the average and median seem to be higher in the mutant, the difference is not statistically significant.

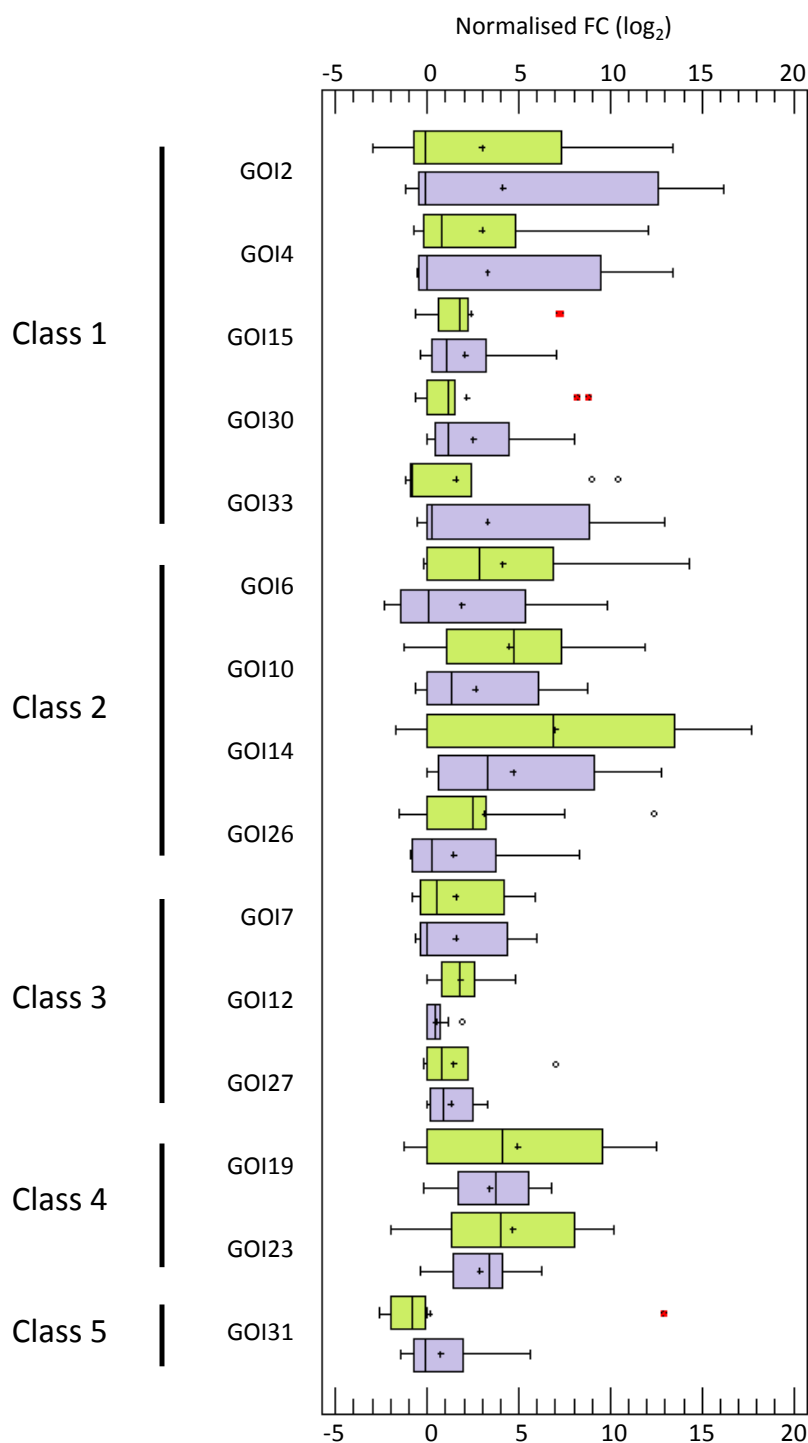


Figure 6.8: Distribution of the expression values for each gene in the 10 conditions tested in *T. versatilis* wild-type and *T. versatilis* $\Delta xlnR$. The transcript values are represented in a log₂ scale. The 5 classes of genes are indicated on the left of the figure. The green boxes correspond to the wild-type strain and the violet boxes, to the $\Delta xlnR$ strain. The left and right boundaries of the box represent the 25th and the 75th percentile, respectively. The line within the box represents the median, the cross represents the

average and the whiskers indicate the highest and the lowest data. The outliers are represented by the circles outside the whiskers. The extreme values are represented with a red asterisk.

This overall distribution of the transcript levels did not give enough information about the influence of the mutation on the behaviour of the genes in the different conditions. We therefore used the scatter plots shown in Figure 6.9 to better visualize the exact influence of the *xlnR* deletion on the expression values of GOI2, GOI10, GOI19 and GOI27, *i.e.* one representative gene of each class. In these scatter plots, a line corresponding to the equation $y = x$ is represented; when a point meets this middle line, it indicates that the transcript level does not change between the two strains.

The transcript levels of the gene GOI2 present a good correlation with an r value of 0.98 between the two strains, but most of the points are placed above the middle line ($y = x$), which indicates that the expression in the mutant strain was higher than in the wild-type strain. When the *xlnR* mutant was exposed to any of the polymers tested (AL18, AL20 or AL22), the expression of GOI2 was significantly higher than in the wild-type. For GOI10 the points in the scatter plot have a correlation coefficient of 0.87, most of the points are below the middle line, indicating that the expression level of this gene in *T. versatilis* $\Delta xlnR$ was lower than in the wild-type. There was a decrease in the transcript levels when the fungus was exposed to polymers, arabinose or xylose, suggesting that XlnR is involved in the induction of this gene in response to the presence of those compounds, in particular xylose (AL38 and AL39). For GOI19, the r value from the scatter plot is 0.92, indicating a good correlation. The specificity of this gene is the clear down regulation for conditions with the higher expression, when the fungus was exposed to xylan (AL20), wheat straw (AL22) or xylose (AL38, AL39). But, the expression remained similar in presence of arabinose or Avicel, suggesting that the transcription of GOI19 in response to these compounds is not controlled by XlnR, whereas in carbon or nitrogen starvation, the expression in the mutant was slightly higher than in the wild-type strain. Altogether, these results perfectly explain the strong shrinkage of the box (Figure 6.8). Finally, the correlation coefficient between the transcript levels in the mutant versus wild-type for GOI27 is very

low: 0.56. In fact, some of the points in the scatter plot are below the middle line and some above, indicating strong overexpression in the mutant strain under some conditions, while other conditions lead to down regulation in the mutant. As a result, the overall distribution appeared unchanged (Figure 6.8), as if the correlation coefficient was close to one. But in this case, closer analysis of the data revealed a huge reorganisation of the transcription in the absence of *XlnR*, which makes difficult to assess what the effect of the deletion of the *xlnR* locus is in the genes from class 3.

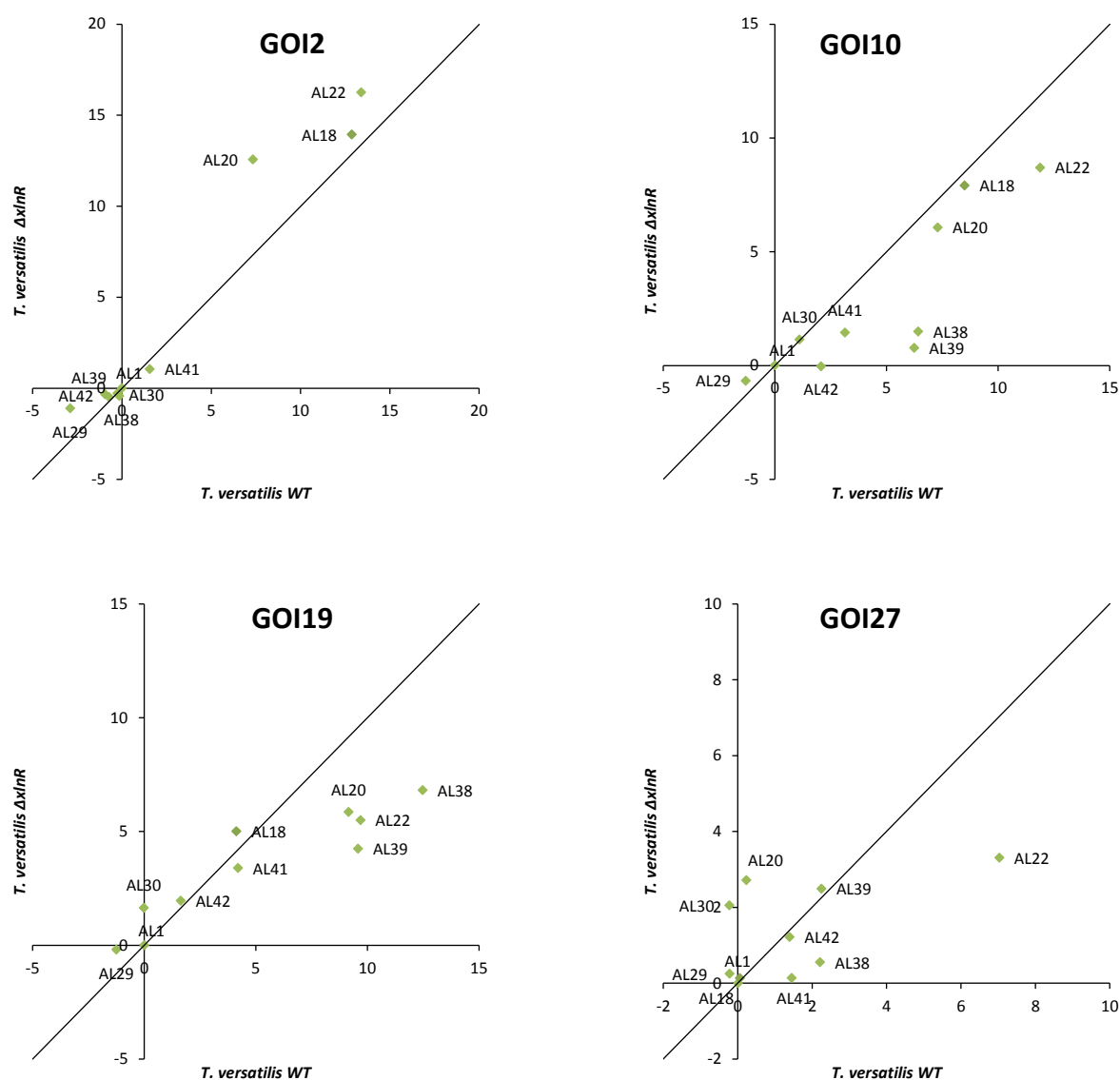


Figure 6.9: Scatter plots of the expression values in the 10 conditions tested in *T. versatilis* wild-type versus *T. versatilis* $\Delta xlnR$. The top left panel shows the GOI2 gene, GOI10 is shown in the top right panel, GOI19 in the bottom left panel and GOI27 in the

bottom right panel. The transcript values are represented in a \log_2 scale. The linear regression shown corresponds to the equation: $y = x$.

In order to visualize which genes are overexpressed or underexpressed in each condition in the absence of *xlnR*, we calculated the ratio between the expression value in the mutant and the expression value in the wild-type strain. The logarithms in base 2 of ratio values are displayed in Figure 6.10 as a heat-map. The ratio is equal to 1 (0 in the heat map), when the deletion of *xlnR* has no influence on the fold change in a condition of interest. The negative values (in red) do not mean a repression of the gene, but only a reduced FC value in the mutant when we compare it to the reference strain. On the contrary, positive values indicate an increase of the FC values (in green) in the mutant strain.

Classes 2 and 4 are composed by the genes that encode proteins that can degrade the hemicellulose, and in most cases there was a decrease in the expression level of these groups of genes when *xlnR* is not present. The loss in the transcript level in the mutant strain was remarkable when the fungus is exposed to xylose (AL38, AL39), wheat straw (AL22) or even cellulose (AL18). Interestingly, the GOI6 and GOI26, which encode the arabinofuranosidase B2 and the endoarabinase respectively, showed an increased transcript level in the mutant strain in presence of xylan (AL20) and GOI19, that encodes a β -xylosidase, presented an increased transcript level in the mutant in presence of cellulose (AL18). The lack of *XlnR* also produced a reduction of the transcript levels in presence of arabinose (AL41, AL42), in particular of the genes from class 2. Although the response observed with arabinose was in general more subtle than the one observed with xylose, there was a strong effect over the GOI6 and GOI26 genes. In presence of arabinose 1%, GOI19 showed an increased transcript level in the mutant.

Concerning the classes 1 and 3, the most remarkable effects were observed with xylan (AL20) and wheat straw (AL22). Surprisingly, the absence of *XlnR* produced an increase in the transcript levels of the genes in the 2 mentioned groups in presence of xylan. Wheat straw, on the other hand

produced an increased transcript level over the 2 cellobiohydrolases, GOI2 and GOI4, and the swollenin, GOI33, from class 1. But the same substrate produced a reduced transcript level in all the other genes from class 1 and class 3, especially GOI12 and GOI27. The absence of XlnR negatively affected the expression of the genes in these 2 classes when the fungus was exposed to Avicel (AL18), arabinose (AL41, AL42) or xylose (AL38, AL39), but this effect was rather moderate. There was mostly a loss of transcripts. However, there were some exceptions: in the mutant, GOI33 showed an enhanced transcript level with Avicel, arabinose or xylose; GOI7 showed a slight increase in transcript level with Avicel and xylose at low concentration (0.2%); GOI27 also showed a very slight increase in the transcript level with Avicel and xylose at high concentration (1%). The hydrophobin that corresponds to the class 5 presented an increased transcript level in all the conditions tested, except wheat straw, where there is an important loss in the transcript level.

The lack of the *xlnR* gene also affected the expression of all the genes studied in carbon (AL30) or nitrogen (AL29) starvation conditions, although the changes observed did not show any clear pattern related to the classes defined previously. We can observe that the genes in classes 4 and 5 had an increased transcript level in the mutant strain in both conditions. GOI2, GOI30 and GOI33 (class 1), GOI14 (class 2), GOI27 (class 3) and also presented enhanced transcript levels in the mutant in these conditions. GOI6 is the only one that showed a lower FC in the carbon starvation condition. All the other genes presented mild changes when the fungus was exposed to these 2 stress conditions.

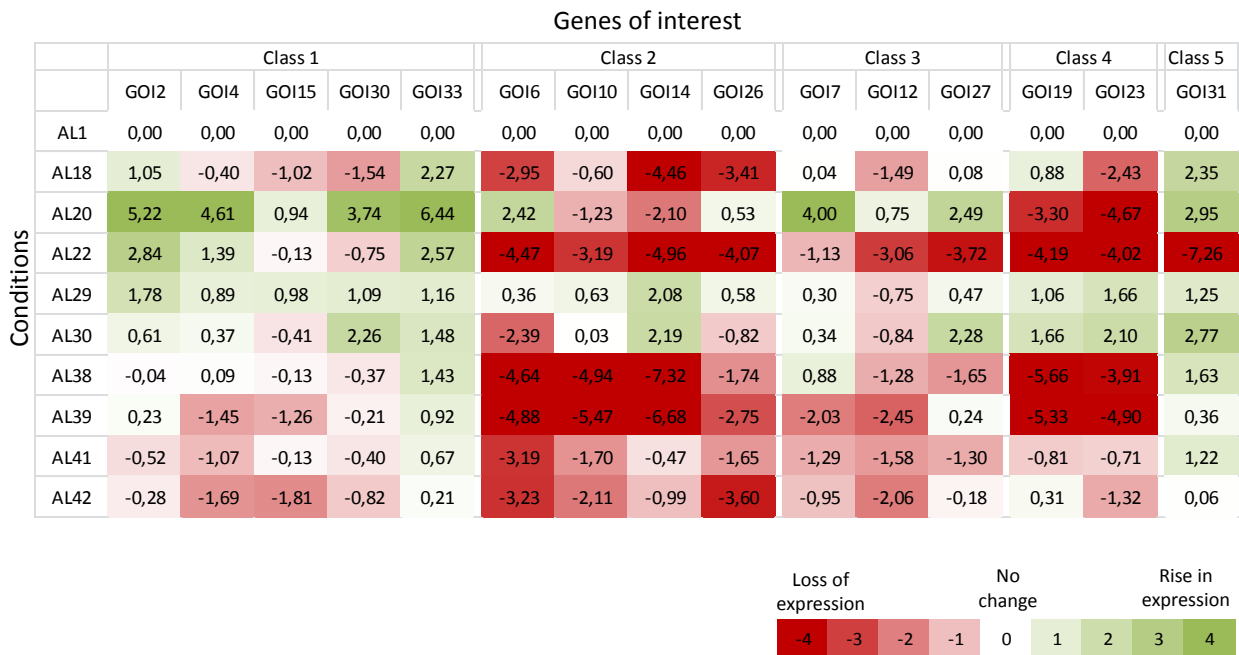


Figure 6.10: Ratio of the expression level of the GOIs in *T. versatilis* $\Delta xlnR$ over *T. versatilis* wild-type. The values shown are the \log_2 of the ratio between the normalised fold changes in the mutant strain over the wild-type strain. Each column corresponds to a gene and each line corresponds to a condition. The negative values (red) indicate a lower expression level in the mutant in comparison to the wild-type strain. Conversely, the positive values (green) indicate an increased expression level in the mutant. And zero (white) indicates that there is no change in the expression levels. The colour scale is shown at the bottom of the figure.

When we look at the expression values of each gene in each condition we can observe fundamental differences with the wild-type strain presented in the previous chapter. As it was shown before, in the absence of XlnR, the genes in class 1 were induced even more strongly when the fungus is exposed to plant polymers, xylan principally. The expression values of the cellobiohydrolase II encoding gene (GOI2) in both strains, the wild-type and the deletion mutant, exposed to different polymers, are shown in Figure 6.11. This gene was strongly induced by all the polymers tested in the wild-type strain, but the induction was enhanced when the xylanolytic activator was missing. The swollenin (GOI33) gene presented an equivalent compartment. The other genes in class 1, *cbhI*

(GOI4), the β -glucosidase (GOI15) and the GH61 (GOI30), presented an induction in response to the presence of xylan in the media that was increased when the *xlnR* regulator was absent.

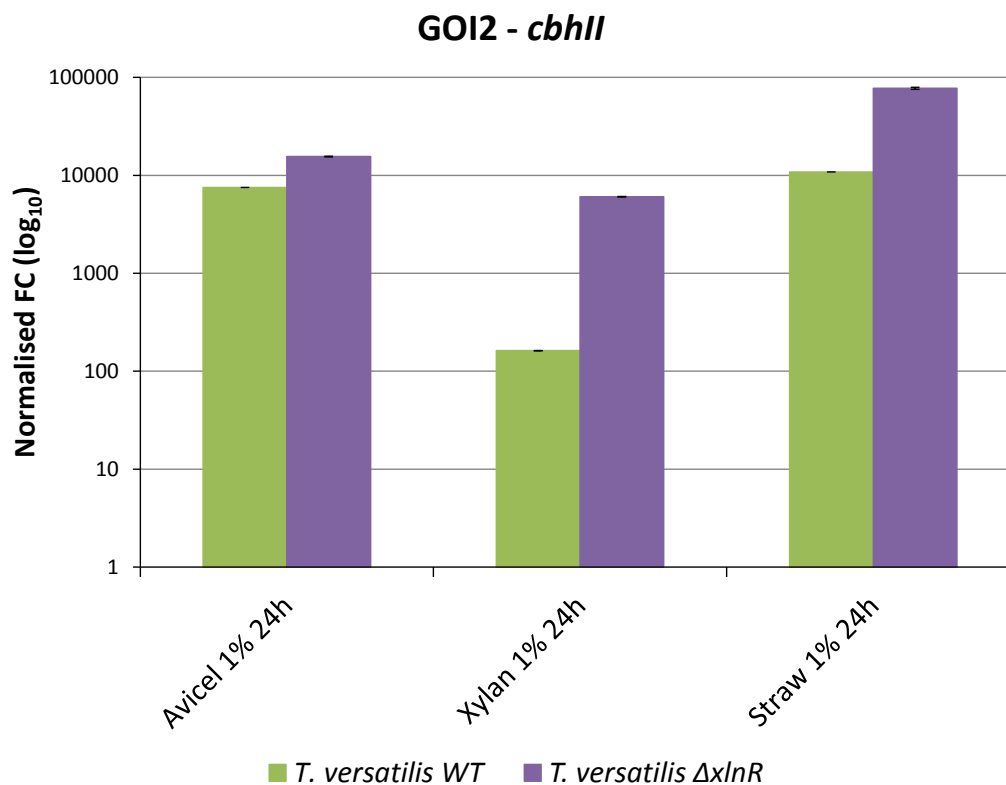


Figure 6.11: Normalised fold change of the GOI2 in response to plant polymers. The transcript levels shown correspond to the strains *T. versatilis* wild-type (green bars) and $\Delta xlnR$ (violet bars). The fold change values were calculated using the glucose condition as a reference and they are represented in a logarithmic scale. Error bars represent SEM.

The two classes that group the genes coding for enzymes that are involved in the degradation of the hemicellulose, class 1 and class 3, suffered an important decrease in the transcript levels in the absence of XlnR. In these conditions, even if the genes were still expressed, the full induction could not be reached. Two examples are shown in Figure 6.12: GOI14, the xylanase D from class 1 and GOI23, the β -xylosidase from class 3. The expression of both genes with any of the polymers or monosaccharides tested was importantly reduced in the mutant in comparison to the wild-type strain. Regarding the monosaccharides, the difference was more dramatic in the presence of xylose

in the media, whereas with arabinose the difference in the transcript levels between the 2 strains was smaller. This is not surprising, since XlnR is an activator that responds to the presence of xylose. The lack of this regulator alters the response to the xylose in the media and the cell has difficulties in reaching a full induction of the genes normally activated in these circumstances. The induction provoked by arabinose is controlled by other transcription factor than XlnR; it is probably AraR the responsible for the up-regulation of these genes. The full induction of these genes when the fungus is confronted to plant polymers is also dependent on XlnR, since in its absence the expression levels dropped significantly. However, there might be other regulators involved in the activation of the genes of classes 1 and 3 that allow their expression in these conditions even when XlnR is missing.

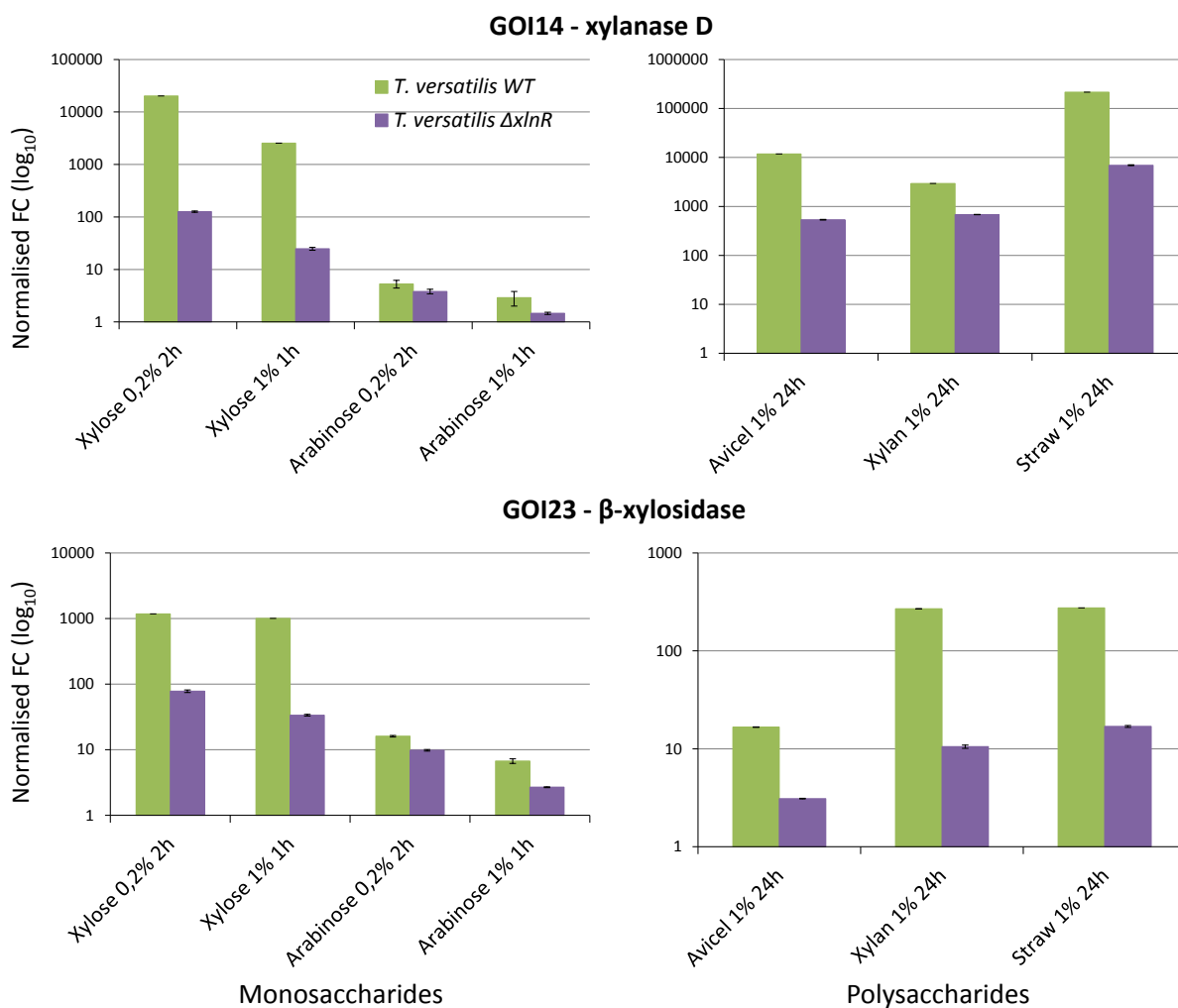


Figure 6.12: Normalised fold change of the GOI14 and GOI23 in response to monosaccharides (left) and plant polymers (right). The transcript levels shown correspond to the strains *T. versatilis* wild-type (green bars) and $\Delta xlnR$ (violet bars). The fold change values were calculated using the glucose condition as a reference and they are represented in a logarithmic scale. Error bars represent SEM.

The hydrophobin (GOI31) that belongs to class 5 presents a particular case. The expression of this gene in *T. versatilis* wild-type and *T. versatilis* $\Delta xlnR$ exposed to plant polymers is presented in Figure 6.13. This gene was repressed by all the substrates tested in the wild-type strain, except the most complex plant polymers. And the situation did not change in the *T. versatilis* $\Delta xlnR$ mutant. Although in the mutant the cellulose (AL18) and the xylan (AL20) produced an induction. The important

activation observed with wheat straw (AL22) in the wild-type, was considerably reduced in the $\Delta xlnR$ strain.

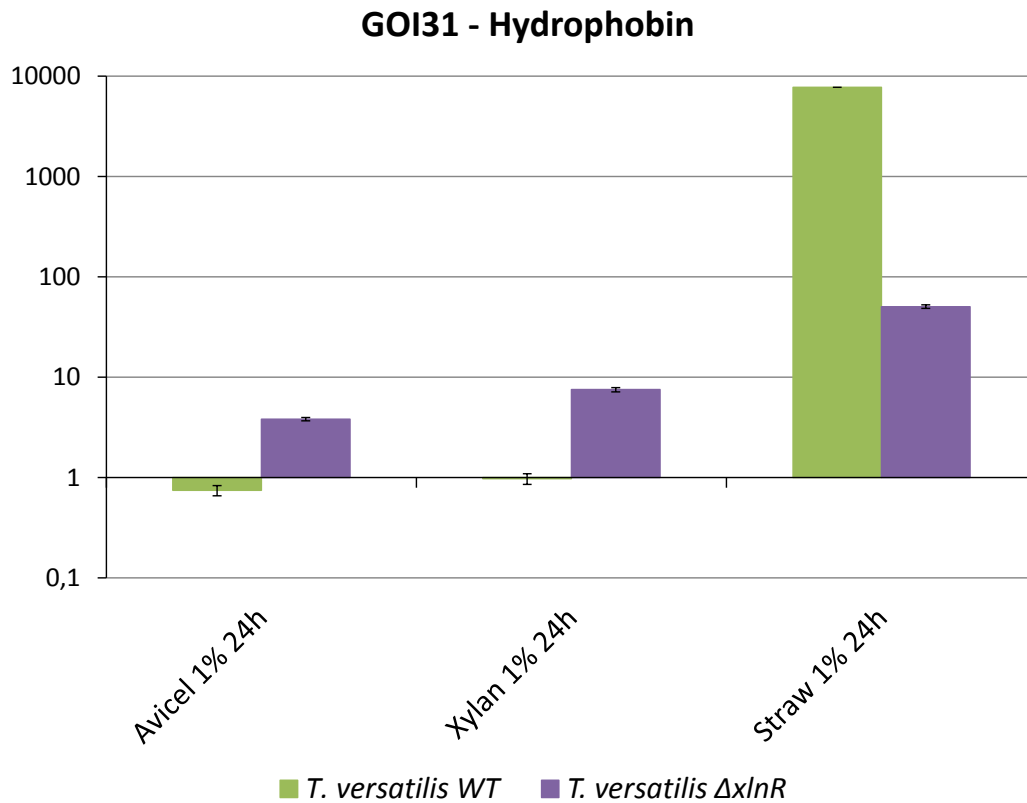


Figure 6.13: Normalised fold change of the GOI31 in response to plant polymers. The transcript levels shown correspond to the strains *T. versatilis* wild-type (green bars) and $\Delta xlnR$ (violet bars). The fold change values were calculated using the glucose condition as a reference and they are represented in a logarithmic scale. Error bars represent SEM.

6.4. Discussion

The genetic analysis in filamentous fungi, and in particular in *T. versatilis*, can be impaired by the lack of tools to produce gene deletion. The limited number of selection markers and the low efficiency of homologous recombination compared to ectopic integration of linear DNA fragments is another concern when trying to genetically modify a fungal strain. A method for gene deletion was developed in *A. niger* and then successfully adapted to *T. versatilis* (Delmas *et al.*, 2014). In order to be able to use this system in *T. versatilis*, it was first necessary to produce a *pyrG* deletion mutant

and to probe that the *pyrG* gene from *A. oryzae* is functional in *T. versatilis* and could be used as a selectable marker later. Once these requirements were fulfilled, the pC3 plasmid was used to successfully delete two different transcription factors in *T. versatilis*. The fact that this system does not require a NHEJ deficient strain ensures that the strain will not be genetically unstable, which is an unwanted consequence of the deletion of the genes involved in the NHEJ (Zhang *et al.*, 2011a). Moreover, the possibility to recycle the selection marker allows the construction of strains with multiple gene deletions, even though this was not tested in *T. versatilis*, in *A. niger* double mutants were successfully obtained. Another important characteristic of the deletion method developed is that it should be effective with other species and genera of filamentous fungi, provided the availability of a *pyrG* deletion strain and a functional complementation of the strain by the *pyrG* gene present in the plasmid. The use of a plasmid importantly improved the transformation efficiency in *T. versatilis*. The integration of the plasmid assures that even with a single clone it is possible to find a deletion mutant after the selective pressure has been released and the excision has occurred. The excision can lead to both, a wild-type locus or a mutant, depending on which side the recombination takes place. In some cases the screening for the deletion mutant can be time-consuming, but there is always the possibility to get a strain that leads to the recombination in the right sense to originate the mutant that was expected. Such was the case of the *T. versatilis* Δ *clrA*, for which the screening was done over more than 50 different clones until the deletion mutants were found. To improve the transformation efficiency and the integration of the plasmid in the correct locus it is important that the fragments flanking the target gene are at least 1 kb length on each side. For *T. versatilis*, the fragments used to construct the cassettes for the different deletion were always about 1.5 kb up and downstream the target gene.

Unfortunately, the production of deletion mutants is not always simple, like in the case of *creA*. As discussed previously, CreA is the major repressor involved in the carbon catabolite repression. The deletion of this gene would result in a strain that is capable of producing hydrolytic enzymes even in

presence of glucose. The construction of this mutant was of particular interest for Adisseo, since a strain that can produce enzymes in an inexpensive substrate can constitute an important benefit. A *creA* deletion mutant would have also been interesting for the study of the transcription regulation of the GH-encoding genes. But this deletion proved to be challenging to obtain, as it was already shown in other filamentous fungi (Shroff *et al.*, 1996, 1997). In fact, in the first attempt the deletion of the full length gene was tried. But the transformants that appeared right after the transformation process were not able to grow in the subsequent passages to fresh media. It is possible that the integrated plasmid disrupted elements upstream or downstream the *creA* locus, preventing the strains to grow afterwards. The second attempt done with a plasmid that was designed to delete only the fragment coding for the zinc finger domain, produced transformants that are able to grow. Unfortunately, the screening for the deletion mutant could not be finished before the end of the thesis works.

The two deletion mutants obtained - $\Delta xlnR$ and $\Delta clrA$ - will allow the study of the regulation of the GH-encoding genes in *T. versatilis*. In fact, preliminary experiments were done with the $\Delta xlnR$ strain. XlnR is a transcriptional regulator that contains a Zn(II)₂Cys₆ binuclear domain and it is very similar to the *A. niger* XlnR; they shares 62% of identity between the amino-acid sequences. The experiments performed with the *T. versatilis* $\Delta xlnR$ clearly showed that XlnR is necessary for the full induction of the hemicellulase genes. This is consistent with the qualitative results obtained with the congo red test, that showed that the *T. versatilis* $\Delta xlnR$ strain has a reduced capacity to degrade the xylan compared to the wild-type strain. In fact, the qPCR results showed that the expression of the genes in classes 2 (the main hemicellulases) and 4 (the β -xylosidases) is seriously affected by the lack of XlnR. However, the genes are still up-regulated in the absence of the xylanolytic regulator. We also observed that the expression of these genes is not importantly affected when the mutant strain is exposed to arabinose. De Souza *et al.* (2013) reported that in *A. niger*, the expression of an α - and a β -galactosidase encoding genes were regulated by both XlnR and AraR. Which leaves the possibility

that the induction observed can be due to the intervention of AraR, besides the de-repression caused by the lack of the carbon catabolite repression.

When we analysed the effect of the lack of XlnR in the transcript levels of the genes in class 1 (the cellulases) we were surprised to see that the genes did not show a reduced transcript level in response to plant polysaccharides. In fact some of them were even more induced than in the wild-type strain. If we take into account that the induction by xylose was very low in the wild-type strain, we realize that these genes are controlled by other transcription factors and that XlnR is not fundamental for their expression, even though it might be involved to some extent. XlnR has been described as a transcription activator of cellulase and hemicellulase genes in different *Aspergilli* (Gielkens *et al.*, 1999; Noguchi *et al.*, 2009) and in *T. reesei* (Stricker *et al.*, 2006). In contrast, Sun *et al.* (2012) revealed that the *N. crassa* homolog, *xln-1*, is essential for the induction of the xylanolytic genes. Despite the fact that it modulates the expression of some cellulase genes, it is not indispensable for their transcription. Similarly, *Fusarium oxysporium* and *Fusarium graminearum* present an analogous behaviour for their XlnR homologs (Brunner *et al.*, 2007; Calero-Nieto *et al.*, 2007). The results obtained in *T. versatilis* concerning the genes in the class 1, which groups genes involved in the cellulose degradation, are consistent with the findings in the last three fungi. XlnR is also involved in the control of the expression of the class 3 hydrolytic enzymes encoding genes. These genes were more induced in the mutant strain than in the wild-type, but only in presence of xylan, whereas with most of the other substrates tested the transcript levels were lower in the mutant. This group is constituted by minor hemicellulases that have very low expression levels in most circumstances. The regulatory mechanisms involved in the up-regulation of this group of genes have not been fully elucidated so far.

The hydrophobin gene shows a very particular expression pattern that is distant from the other genes studied. But, the strong induction of the hydrophobin gene observed in the wild-type strain in presence of lignocellulosic material is importantly reduced in the *xlnR* deletion mutant, suggesting

that XlnR is in fact involved in the activation of the transcription of this gene, though it might not be the only regulator implicated. This gene may have a very different regulatory mechanism controlling its expression and further experiments are necessary to depict it.

The Clr-1/ClrA transcription factor was first described in *N. crassa* and in *A. nidulans*. It was shown that this regulator is essential for the expression of cellulase genes in response to the presence of cellobiose (Coradetti *et al.*, 2012; Znameroski *et al.*, 2012). In fact, when we tried to induce the expression of the GH-encoding genes by exposing the *T. versatilis* Δ *clrA* to wheat straw, which typically is a very strong inducer, no cellulase activity was observed in the congo red test. More interestingly, no xylanase activity was observed either. This indicates that the ClrA regulator has an important function in the activation of the cellulase and xylanase encoding genes. XlnR is also involved in the expression of these two groups of genes, but principally in the expression of xylanases, as it has been shown previously through the qPCR results and the congo red test. Yet, the effect of the deletion of *xlnR* is not as dramatic as the deletion of *clrA*. The *T. versatilis* ClrA homolog seems to have a very wide and important role. RT-qPCR experiments will allow determining to what extent the lack of ClrA negatively affects the expression of the GH genes.

7. General conclusions and perspectives

7.1. Conclusions

The main objective of this thesis was to characterize the regulatory networks that control the expression of the genes encoding for enzymes involved in the degradation of the plant polysaccharides. We analysed the growth of *T. versatilis* with different carbon sources. Then we used a genomic approach, to identify the genes involved in the production and secretion of hydrolytic enzymes, and a transcriptomic approaches to study the expression pattern of the genes coding for hydrolytic enzymes.

***T. versatilis* shows a good growth rate on complex carbon sources and prefers ammonium over nitrate as nitrogen source.** The carbon metabolism, that serves for the generation of energy and growth are closely linked processes and fungi have evolved means to make use of the sugars present in the environment, transport them into the cell and then metabolise them to get energy. The easily metabolisable simple sugars (monomers) provide the energy for enhanced hyphal growth. However, *T. versatilis* showed a very slow growth rate with most of the sugars tested on solid media. And we were surprised to observe that this fungus has no difficulties to grow with complex polysaccharides derived from the plant cell wall. This indicates that the fungus has all the necessary tools to degrade such complex substrates, which is consistent with the mutagenesis process that led to the selection of this industrial strain, where enzymes-overproducing strains were selected on complex polymers. The comparison of the *T. versatilis* mycelia with other fungi – *T. reesei*, *A. niger* and *A. nidulans* – growing on plates with complex lignocellulosic substrate showed that *T. versatilis* produces a thicker and covering mycelium. This suggests that *T. versatilis* may be capable of branching more frequently to cover the agar surface and therefore secrete the enzymes that degrade the substrate efficiently allowing the fungus to colonise this kind of complex media. We also observed the influence of the

nitrogen source on the growth rate of *T. versatilis*, who has preference for the ammonium over the nitrate, the second one being more difficult to metabolize.

The analysis of the genome of *T. versatilis* confirmed that this fungus has the genetic capacity to cope with the saprophytic lifestyle in environments with complex sugars that need to be degraded to monomers before being taken up and metabolized. However, the genome of *T. versatilis* available at the time of the analysis was not ideal, since the annotation was particularly problematic. The errors in the automatic annotation needed to be manually curated in order to avoid possible mistakes in the prediction of gene functions and the development of experimental approaches based on the gene sequences. The manual curation carried out over the three groups of genes involved in the production and secretion of enzymes for the degradation of the plant polymers – hydrolytic enzymes encoding genes, transcriptional regulators controlling their expression and secretion pathway genes – allowed us to detect an important panel of genes in *T. versatilis*. The high homology of these genes with the fungi *T. marneffeii* and *T. stipitatus* was remarkable. The close phylogenetic relation between the three of them was clearly shown, not only by the high level of sequence homology, but also by the syntenic regions found when comparing the genes location.

Within the limited number of GH families analysed in the *T. versatilis* genome, we detected a large number and variety of GH-encoding genes that covered all the enzymatic activities necessary for the degradation of the different polymers, including cellulose, hemicellulose and pectins. When comparing the panel of GH-encoding genes of *T. versatilis* with other fungi, we confirmed that this fungus is amongst the fungi with the highest number of genes coding for hydrolases. The transcription factors that are known to control the expression of the GH genes were all found in the *T. versatilis* genome, with only one exception, ACEII, an activator found in *T. reesei* that is missing in several filamentous fungi. The gene coding for the repressor ACEI, on the other hand, was present but its sequence was truncated, suggesting that the protein produced may not be functional as a repressor. The genes involved in the secretion pathway were also analysed, and we observed that

they have a high level of homology with *A. nidulans*, which was the fungus used as model for this analysis. The high similarities indicate that the secretion pathway is very well conserved among fungi and that this is an essential element for the fungal lifestyle.

We also detected an important number of genes coding for hydrophobic proteins, and swollenin-encoding genes were also present. These might be related with the degradation of the plant polymers, although their implication in this process has not been demonstrated yet. However, we observed that the expression of many of these genes was up-regulated in presence of lignocellulosic substrates, suggesting that their promoters are activated in response to that kind of carbon source, although we could not determine which transcription factor is involved.

The transcriptomic study of *T. versatilis* revealed the enormous reorganisation of the transcriptome when the fungus is exposed to glucose and then transferred to lignocellulose. The necessary adaptation of the fungus to such different conditions was shown in this experiment; a large number of genes were significantly differently expressed and a group of genes switched from antisense to sense transcription, or vice versa, all this showing the massive transcriptional switch made by the fungus to cope with the change in its environment. Amongst the up-regulated genes in lignocellulose, there were a large number of GH genes, swollenins, hydrophobins, sugar transporters and esterases. The reduced group of genes constituted by the GH-encoding genes produced 13% of the transcriptome, which is a reflection of the investment made by the fungus in order to degrade the lignocellulose present in the medium. The up-regulated genes also included transcription factors, mainly Zn(II)₂Cys₆ binuclear domain transcription factors. *xlnR* and *clrA* were included in this group, but there were 11 other unknown transcription factors. They constitute a hint that might be worth exploring, since they are probably involved in the strong activation of the GH-encoding genes and could be part of that regulatory network. As mentioned before, this study also allowed to identify the presence of NATs for some of the *T. versatilis* genes. The NATs might be involved in particular post-

transcriptional mechanism of regulation. Further experiments would be necessary to determine the specific role of each one of these antisense RNAs.

The transcriptomic results allowed the selection of new putative reference genes for RT-qPCR data normalisation in filamentous fungi. To deeply explore the expression of the GH-encoding genes we decided to use the RT-qPCR technique, which is perfectly suitable to study a limited number of genes in a large variety of conditions. Before starting to analyse the transcript levels of the genes of interest, it was necessary to find good reference genes that could be used to normalise the expression data. The RNA-seq results clearly showed that the “classis” housekeeping genes were differently expressed, therefore using them for RT-qPCR data normalisation can lead to wrong conclusions. The RNA-seq results were used to select a group of genes with different biological functions that were stably expressed. Their stability was assessed in a wide range of conditions and they were classified using the geNorm algorithm. Three genes, *ubcB*, *sac7* and *psm1*, were certified as the best reference genes for accurate normalisation of expression levels, but the overall good stability of the 12 candidates led us to show that any combination of 3 of them could lead to minimal normalisation bias. The suitability of these reference genes to be used for data normalisation in other filamentous fungi besides *T. versatilis* was studied. Using transcriptomic data from phylogenetically distant fungi we could classify the genes according to their stability. A group of very promising genes was defined and it included the *ubcB* (ubiquitin carrier protein), *sac7* (Rho GTPase activator), *fis1* (mitochondrial membrane fission protein), *sarA* (secretion associated GTP-binding protein), and two genes homologous to *S. cerevisiae* *TFC1* and *UBC6*. The first 3 genes in this group correspond to the genes selected on *T. versatilis*. There is no universal gene stable in any sample and/or organism of interest. It seems therefore utopic to think escaping from sporadic differential expression, which strengthens the need for systematic validation of the stability of the reference genes and the use of the geometric mean of at least 3 genes to normalise the data and get reliable results.

Different patterns in the expression of the genes coding for cellulase and hemicellulase were revealed in response to various inducers. A group of genes of interest that included cellulases, hemicellulases and other auxiliary proteins encoding genes were studied by RT-qPCR and the previously validated reference genes were used to normalise the expression values. We analysed the expression of all these genes when the fungus is exposed to a variety of conditions that included carbon sources and stress conditions. The genes were classified according to the correlation between the expression levels in each of the conditions tested. This analysis allowed to determine 5 different classes of genes. Since the behaviour of the genes in the same class is very similar to each other, it is possible to foresee that the genes share the same regulatory mechanisms and that their promoter regions have similar elements. An analysis of the promoters can be considered to predict which genes are controlled by which transcription factor. Independently of the classes, we observed that all the genes studied are strongly induced by the lignocellulose, which is consistent with the RNA-seq results. Xylose also proved to be a very good inducer, better than arabinose. But the use of high concentrations of any of the monosaccharides caused a similar effect than the one observed when the carbon catabolite repression induced by the presence of glucose is activated. Cellobiose and xylobiose were also strong inducers of the cellulase and hemicellulase genes, respectively, confirming that the products of the degradation of the plant polymers are the real inducers of the hydrolase genes. These substrates probably act through the action of the ClrA and XlnR activators, but further experiments would be necessary to confirm this theory. The hydrophobin gene was only activated by the presence of lignocellulosic material, but none of the isolated inducers tested provoked any response. The induction mechanism of the expression of this gene could not be elucidated yet. The results of expression of the GOIs when the fungus is exposed to stress conditions suggested that some of genes, like the cellobiohydrolase genes, the *abf-B2*, the *xynD* and the *xynC*, are under the control of the PacC regulator, which activates their expression in alkaline pH. We also observed a de-repression of some of the GOIs when no carbon is supplemented in the medium, which may be related to the release of the carbon catabolite repression exerted by CreA.

A transformation method with a recyclable selectable marker gene was adapted to *T. versatilis* and used for the construction of deletion mutants. In order to analyse the regulatory mechanisms controlling the GH-encoding genes it was first necessary to create the genetic engineering toolbox to produce deletion mutants in *T. versatilis*. We adapted to our fungus a transformation method, first designed for *A. niger*. This method allows the recycling of the selectable marker gene, which in this case was the *pyrG*, and it works in NHEJ-proficient strains. Moreover, the use of a circular plasmid improved the transformation efficiency in *T. versatilis*. The only pre-requisite was the construction of a *pyrG* deletion mutant, which was achieved by using a linear deletion cassette. With this method we were able to construct deletion mutants in two different transcription factors: *xlnR* and *clrA*.

The *T. versatilis* $\Delta xlnR$ strain showed a reduced expression of the hemicellulase genes, but the expression of the cellulase genes was not affected. Both RT-qPCR experiments and the congo red test showed that the *T. versatilis* $\Delta xlnR$ strain has a reduced production of hemicellulase genes, although the expression of these genes was not completely abolished. The lack of the xylanolytic regulator did not affect the expression of the cellulase genes. Some of these genes were even more induced, suggesting that XlnR is not essential for the expression of the cellulase genes, but it is involved in the expression of the hemicellulase genes and the hydrophobin gene. The congo red test on the *T. versatilis* $\Delta clrA$ was much more determining; no hemicellulase or cellulase activity was observed after induction of the enzymes on medium containing wheat straw. These results suggest that ClrA is implicated in the induction of both the cellulase and the hemicellulase genes, and that it is an essential transcription factor in this network.

7.2. Perspectives

Further studies to characterise the roles of the different transcription factors during the process of carbohydrate degradation would focus on the following:

- The RNA-seq data obtained was recently re-mapped against the genome with the new annotation. This new mapping gave a better quality transcriptomic data. The data obtained has potentially a lot of information and an analysis in depth should be performed in order to identify the groups of genes differentially expressed in the three conditions tested and extract the pieces of information that will help to understand the response of the fungus to these conditions.
- With the tools already produced for the study of gene expression through RT-qPCR it will be possible to carry out an exhaustive analysis of the transcript level of the GH-encoding genes. The expression of candidates, such as cellobiohydrolases, arabinofuranosidases, xylanases and other auxiliary proteins will be analysed in response to different carbon sources and stress conditions, and in the two deletion mutants obtained during the thesis. The use of the Fluidigm technology can therefore be considered, given the large number of samples and genes that will be analysed and the requirement of multiple biological replicates for robust conclusions.
- Additionally, the analysis of the transcript levels of the putative transcriptional factors, revealed by the high throughput transcriptomic study, is worth to be considered, since it could give us a hint on the implications of these factors in the enzyme production process.
- The new and improved annotation, together with the clustering results that shall be obtained from this high-throughput fluidigm-based analysis of gene expression, it would also be possible and relevant to carry on an *in-silico* analysis if the promoter region of the genes of interest, to detect possible binding sites for the transcriptional factors, and in this way predict how the genes might be regulated.

Appendix I

Table AI: Number of genes encoding for different GH enzymes and its corresponding Pfam domain numbers in 7 filamentous fungi.

Enzyme	CAZy family	Pfam	<i>T. versatilis</i>	<i>A. niger</i>	<i>A. nidulans</i>	<i>A. fumigatus</i>	<i>A. oryzae</i>	<i>P. Chrysogenum</i>	<i>T. reesei</i>
β -glucosidase	GH1	PF00232	4	3	3	5	3	3	2
β -galactosidase	GH2	PF00703	2	0	3	2	1	2	1
β -mannosidase			6	3	3	2	3	2	6
β -glucuronidase			1	3	3	1	2	1	0
Exo- β -d-glucosaminidase			1	0	0	2	1	1	1
β -glucosidase	GH3	PF00933	20	12	15	12	16	12	8
β -xylosidase			2	2	3	2	5	1	2
β -N-acetylhexosaminidase			2	3	2	3	2	3	2
β -1,4-endoglucanase	GH5	PF00150	4	4	4	6	4	3	4
β -1,4-endomannanase			2	1	6	2	2	3	1
Exo-1,3- β -glucanase			2	4	3	2	3	1	1
Endo-1,6- β -glucosidase			1	0	1	1	2	1	0
Endoglycoceramidase			0	1	1	2	2	0	0
Cellobiohydrolase (non reducing end)	GH6	PF01341	1	2	2	1	1	1	1
Cellobiohydrolase (reducing end)	GH7	PF00840	2	2	2	2	2	3	1
β -1,4-endoglucanase			1	0	1	2	1	0	1
β -1,4-endoxylanase	GH10	PF00331	1	1	3	4	4	3	1
β -1,4-endoxylanase	GH11	PF00457	7	4	2	3	4	1	3
Xyloglucan-active β -1,4-endoglucanase	GH12	PF01670	2	3	1	4	4	2	2
α -amylase	GH13	PF00128	8	4	3	2	3	7	1
α -1,6-glucosidase			2	2	2	5	5	1	1
α -1,3-glucan synthase			2	5	2	3	3	4	0
GPI-anchored α -amylase like			0	3	3	3	2	0	0
Intracellular α -amylase like			0	2	1	1	2	0	0
Glycogen debranching enzyme			1	2	2	1	2	0	1
Glucoamylase	GH15	PF00723	3	2	2	5	3	2	2
Endomannanase	GH26	PF02156	1	1	3	0	1	1	0

Appendix

α -galactosidase	GH27	PF02065	5	5	3	5	3	1	8
Endopolygalacturonase	GH28	PF00295	6	7	3	4	4	1	1
Endorhamnogalacturonase			5	6	1	2	6	0	0
Exopolygalacturonase			4	4	3	3	2	3	2
Exorhamnogalacturonase			0	3	1	2	4	0	0
Xylogalacturonan hydrolase			1	1	1	1	2	1	1
α -l-fucosidase	GH29	PF01120	2	1	0	0	0	0	0
α -glucosidase	GH31	PF01055	10	7	9	6	10	8	3
α -xylosidase			1	0	1	1	1	0	1
Endoinulinase	GH32	PF00251	1	1	0	1	0	0	0
Exoinulinase			2	1	1	1	0	2	0
Invertase			4	3	1	2	4	5	0
β -galactosidase	GH35	PF01301	5	5	4	5	6	2	1
α -galactosidase	GH36		4	2	4	4	3	1	1
α -l-arabinofuranosidase	GH43	PF04616	1	2	3	3	3	0	1
Endoarabinase			4	4	4	6	6	7	0
β -xylosidase			5	2	5	2	5	4	1
β -1,4-xylosidase/ α -arabinofuranosidase			3	1	1	4	3	1	0
Galactan 1,3- β -galactosidase			0	1	2	1	3	1	0
β -1,4-endoglucanase	GH45	PF02015	2	0	1	1	0	0	1
α -l-arabinofuranosidase	GH51		2	3	2	2	3	3	0
Endo-1,4- β -galactanase	GH53	PF07745	1	2	1	1	1	1	0
α -l-arabinofuranosidase	GH54		5	1	1	1	1	1	2
β -1,4-endoglucanase	GH61	PF03443	2	7	9	7	8	4	3
Arabinoxylan arabinofuranohydrolase	GH62	PF03664	3	1	2	2	2	1	1
α -glucuronidase	GH67	PF03648	3	1	1	1	1	1	1
Endoglucanase	GH74		1	1	2	2	0	0	1
α -l-rhamnosidase	GH78	PF05592	8	8	8	6	8	5	1
d-4,5 unsaturated glucuronyl hydrolase	GH88	PF07470	2	1	2	2	3	5	1
Exoarabinanase	GH93		0	0	2	3	3	1	0
α -l-fucosidase	GH95		0	2	2	2	3	1	4
Total			170	152	156	161	181	118	77

Appendix II

Tracking the best reference genes for RT-qPCR data normalization in filamentous fungi

Agustina Llanos^{1,2,3,4}, Jean Marie François^{1,2,3} and Jean-Luc Parrou^{1,2,3,§}

Author affiliations :

¹Université de Toulouse; INSA, UPS, INP; LISBP, 135 Avenue de Rangueil, F-31077 Toulouse, France ;

²INRA, UMR792 Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse, France ;

³CNRS, UMR5504, F-31400 Toulouse, France

⁴Cinabio-Adisseo France S.A.S., 135 Avenue de Rangueil, 31077 Toulouse, France

Email addresses: llanos@insa-toulouse.fr

fran_jm@insa-toulouse.fr

parrou@insa-toulouse.fr

[§]Corresponding author: parrou@insa-toulouse.fr ; Tel.: 33-561-559-423; Fax: 33-561-559-400;

<http://www.lisbp.fr/fr/index.html>;

Abstract

Background: A critical step in the RT-qPCR workflow for studying gene expression is data normalization, one of the strategies relying on the availability of reliable reference genes. This study aimed to identify and validate a selection of reference genes for relative quantification in *Talaromyces versatilis*, a relevant industrial filamentous fungus for the production of an enzymatic cocktail. Beyond *T. versatilis*, this study also aimed to propose reference genes that are applicable more widely for RT-qPCR data normalization in filamentous fungi.

Results: A selection of stable, potential reference genes was carried out *in silico* from RNA-seq based transcriptomic data obtained from *T. versatilis* exposed to wheat straw. A dozen functionally unrelated candidate genes were analysed by RT-qPCR assays over more than 30 relevant culture conditions. By using geNorm, we showed that most of these candidate genes had stable transcript level in most of the conditions, including various stresses, a wide variety of carbon sources and during conidial germination. The overall robustness of these genes was explored further by showing that any combination and geometric averaging of 3 of them led to minimal normalization bias in any condition. To extend the relevance of the study beyond *T. versatilis*, we challenged the stability of these reference genes together with sixteen other classically used genes such as β -tubulin, actin or glyceraldehyde-3-phosphate dehydrogenase, in a representative sample of about 100 RNA-seq datasets. These freely available datasets were obtained from 18 phylogenetically distant filamentous fungi exposed to prevalent experimental conditions. Although this wide analysis demonstrated that each of the chosen genes exhibited sporadic up- or down-regulation in a few conditions, a hierarchical clustering of all these genes allowed the identification of a promising group of 6 genes, which presented weak expression changes and no tendency to up- or down-regulation over the whole set of conditions. This group included *ubcB*, *sac7* and *fis1* genes, three genes suitable for normalization in *T. versatilis*, as well as *sarA* and two genes homologous to *S. cerevisiae* *TFC1* and *UBC6* that were previously validated for their use in this yeast species.

Conclusions: We propose a set of 6 genes that can be used as reference genes in RT-qPCR data normalization in any field of fungal biology. However, we recommend that the uniform transcription of these genes is tested by systematic experimental validation and to use the geometric averaging of at least 3 of the best ones. This will minimize the bias in normalization and will support trustworthy biological conclusions.

Keywords

Filamentous fungi, *Talaromyces*, RNA-seq, RT-qPCR, reference genes, normalization, gene expression.

Background

Filamentous fungi are involved in several natural and industrial processes. They have long been used for the production of additives used in food and beverages (Archer *et al.*, 2008). Some fungi produce enzymes that degrade lignocellulosic material with applications in food, feed, textile, pulp and paper industries (Floudas *et al.*, 2012; de Vries, 2003). The genera *Penicillium* and *Aspergillus* are the most biotechnologically important fungi, due to their ability to produce secondary metabolites, organic acids or enzymes, but recent genome sequences of hundreds of fungal species indicate that the potential of fungi has been substantially underestimated (Brakhage & Schroeckh, 2011; Wiemann & Keller, 2014). *Talaromyces* is another industrially relevant genus closely related to *Penicillium* (Houbraken *et al.*, 2014), among which *Talaromyces versatilis* is exploited for the production of a commercial cocktail called “Rovabio Excel™” that is used as feed additive for enhancing digestibility of cereal based diets. However, fungi are not restricted to biotechnologically relevant organisms. Recent estimates suggest that more than 5 million fungal species exist in this monophyletic kingdom, the huge majority being in the Ascomycota and Basidiomycota phyla (Blackwell, 2011). Fungi have considerable impact in agriculture, as fungi are capable of intimate symbiotic associations with plants as in the case of *Rhizophagus irregularis* (Tisserant *et al.*, 2013) while some species are economically

Appendix

serious plant pathogens (Schmidt & Panstruga, 2011), e.g. *Leptosphaeria maculans* (Rouxel *et al.*, 2011), *Blumeria graminis* (Hacquard *et al.*, 2013), *Rhizoctonia solani* (Schmidt & Panstruga, 2011; Zheng *et al.*, 2013), *Magnaporthe grisea* (Fernandez & Wilson, 2014). Finally, they not only draw interest as pathogens of invertebrate animals, but they are also harmful for human health, as for example with the production of mycotoxins and allergens (Thornton & Wills, 2013), with several species, including *Aspergillus fumigatus*, causing invasive disease (Dagenais & Keller, 2009). For all these reasons, and aided by extraordinary advances in genome sequencing facilities (Quail *et al.*, 2012; Zhang *et al.*, 2011b), there has been a tremendous effort to pursue the sequencing of filamentous fungi. The availability of genomic sequences from several fungi has favoured the rapid development of high throughput transcriptomic studies and functional genomics analysis.

Better knowledge of gene function usually begins by investigating expression of the genes of interest (GOIs) under a broad set of culture conditions. Several techniques have been developed over the years to measure expression levels, among which the coupling between reverse transcription and quantitative (real-time) PCR (RT-qPCR) appears to be the most appropriate to study limited numbers of genes in large sets of conditions (Bustin *et al.*, 2005; Wong & Medrano, 2005). Significant technical advances made this mRNA quantification method very accessible, highly specific and sensitive, but numerous critical issues remain that limit the ability to draw meaningful conclusions (Derveaux *et al.*, 2010). The Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines help in the design of experiments, to keep track of the experimental data and to improve analysis (Bustin *et al.*, 2009; Huggett & Bustin, 2011). Importantly, and no matter what the technique for measuring gene expression is, data normalization is a critical step. Performance and pitfalls of the different normalisation strategies has already been compared in a number of dedicated review articles (Huggett *et al.*, 2005; Wong & Medrano, 2005). Few articles promote the use of external controls (Devonshire *et al.*, 2010; Liu & Slininger, 2007), a normalization strategy stimulated by the ERCC or EQUAL-quant programs (Baker *et al.*, 2005; Ramsden *et al.*, 2006), and especially relevant for the assessment of technical robustness in clinical and biological diagnostic laboratories.

But normalization of gene expression levels by reference genes (internal controls) is most certainly the gold standard, even if it is now clearly established that the use of a single gene is not acceptable, as there is not a single gene that has a stable transcript level over all kinds of culture conditions or among different cell types (Hruz *et al.*, 2011; Kozera & Rapacz, 2013; Thellin *et al.*, 2009; Vandesompele *et al.*, 2009). The main challenge concerning these internal controls is the circular problem in evaluating expression stability of a candidate normalization gene (Andersen *et al.*, 2004), *i.e.* how can the expression stability of a candidate be evaluated if no reliable measure is available to normalize the candidate? To overcome this circular problem, Vandesompele *et al.* (Vandesompele *et al.*, 2002) first developed more than ten years ago a method called geNorm, which allows the evaluation and ranking of candidate reference genes in terms of expression stability (or suitability as normalizing gene). In a subsequent step, the algorithm is able to indicate how many reference genes are optimally required to remove most of the technical variation. Other algorithms were then developed (*e.g.* Normfinder (Andersen *et al.*, 2004) or BestKeeper (Pfaffl *et al.*, 2004)) and were presented in a comprehensive survey (Vandesompele *et al.*, 2009). Good practice in data normalization for gene expression analysis therefore relies on the identification, experimental validation and use of several reference genes. In filamentous fungi, such efforts have been observed with recent publications dedicated to the validation of suitable reference genes under specific experimental contexts (Bohle *et al.*, 2007; Fang & Bidochka, 2006; Hacquard *et al.*, 2011; Huang Xueling *et al.*, 2012; Raffaello & Asiegbu, 2013; Steiger *et al.*, 2010; Vieira *et al.*, 2011; Yan & Liou, 2006; Zampieri *et al.*, 2014; Zhang *et al.*, 2009; Zhou *et al.*, 2012). In Zhou's work (Zhou *et al.*, 2012), *cypB* and *crzA* were evaluated because of their stability in transcriptomic datasets. Similarly, Kim and Yun (Kim & Yun, 2011) selected 8 reference genes from transcriptomic data available with *Fusarium graminearum*. Such an approach was an exception, as most often, authors have evaluated more classic "housekeeping genes" encoding for example actin, glyceraldehyde-3-phosphate dehydrogenase or β -tubulin, which are still and too frequently used as single, non-validated reference genes.

Appendix

During the course of the RNA-seq based transcriptomic analysis of the industrial strain *T. versatilis* exposed to wheat straw, it was found that most of the classical reference genes exhibited expression changes in the presence of this lignocellulosic substrate (unpublished data). This finding prompted the formulation of a list of putative reference genes and validation of their expression stability in *T. versatilis* cultivated under more than 30 different relevant conditions, following the MIQE guidelines for robust and reliable RT-qPCR expression data acquisition and treatment. Finally, 94 RNA-seq based transcriptomic datasets from 18 phylogenetically distant filamentous fungi were scrutinized, including datasets from industrially important or model species as well as plant or animal interacting fungi, to demonstrate that some of the candidate genes suitable for *T. versatilis* can be proposed as promising reference genes for data normalization in RT-qPCR analysis in other filamentous fungi.

Methods

Strain and culture conditions

The industrial strain used in this work, *Talaromyces versatilis* (basionyme *Penicillium funiculosum*, IMI378536), is an ADISSEO proprietary strain (patent no. W0 99/57325). Spores of *T. versatilis* were obtained by growing the strain on Potato Dextrose Agar (PDA) plates and the spores were used to inoculate liquid medium. The minimal medium (MM) contained for 1 L: 1.9 g KH_2PO_4 , 0.65 g KCl, 0.65 g MgSO_4 , 12.5 mg ZnSO_4 , 12.5 mg MnCl_2 , 12.5 mg FeSO_4 , 5 g NH_4Cl . The MM was supplemented with 10 g/L glucose as the sole carbon source, unless otherwise stated. The pH was adjusted to 6.0 with 50 mM KH_2PO_4 . The liquid medium was inoculated with 2×10^5 spores/mL in Erlenmeyer flasks. The cultures were carried out at 30 °C and agitated at 150 rpm for 48 h.

Mycelia samples

A summary table of the culture conditions is presented as Additional file 1. To prepare mycelia samples of *T. versatilis* exposed to different carbon sources, the mycelia were grown for 48 h in MM

broth culture and were filtered through Miracloth (Merck), washed with MM without carbon source and transferred to fresh media containing the desired carbon sources. The cultures were incubated from 30 minutes to 2 hours for growth on monosaccharides (arabinose 0.2% (w/v) or xylose 0.2%) or disaccharides (cellobiose 0.2%, xylobiose 0.2% or thio-gentiobiose 0.2%), or up to 24 hours for the cultures containing complex carbon sources (Avicel 1%, Arbocel 1%, beechwood xylan 1%, ball-milled wheat straw 1% or micronized wheat bran 1%). For exposure to stress, the *T. versatilis* mycelia grown for 48 h in MM medium were filtered through Miracloth, washed with MM without carbon source and transferred to MM supplemented with 0.5 M KCl for salt stress, MM without glucose for carbon starvation, MM without ammonium for nitrogen starvation, and MM with the pH adjusted to 2 or 8 for pH stress. The cultures were incubated at 30 °C for 1 h before sampling. For the temperature stress, the mycelia were similarly collected and transferred to a pre-heated MM broth culture for an additional 1 h at 40 °C. Samples of about 50 mg of mycelium were then collected by filtration through Miracloth and flash frozen in liquid nitrogen.

The conidia at different developmental stages were prepared after inoculating MM with 2×10^5 spores/mL. Samples were harvested by centrifugation at 3000 g for 2 min, after incubation of the spores at 30 °C, 150 rpm for 2 h (no morphological change), 4 h (early swelling), 8 h (late swelling), 12 h (germ tube on one side of the conidia) and 16 h (hyphae already visible). 500 µL of pre-heated RNA extraction buffer (NaCl 0.6 M, sodium acetate 0.2 M, EDTA 0.1 M, SDS 4%) were added to each sample before immersing in liquid nitrogen.

RNA extraction and cDNA synthesis

Mycelia and conidia samples were mechanically disrupted using the TissueLyser II (Quiagen). Frozen mycelia samples were disrupted with a single 5 mm stainless steel bead (Quiagen), whereas thawed conidia preparations were mixed with *approx.* 150 µL of 625 µm glass beads (Sigma). Both were submitted to two high-speed shaking cycles of 3 minutes at 30 Hz. Total RNA was isolated from disrupted mycelia samples using the GeneJET Plant RNA Purification Mini Kit (Thermo). An on-column

Appendix

DNase I treatment (Thermo – Reference #EN0521) was added to the protocol, applying 100 μL of the DNase I mix (50 μL of DNase I, 10 μL of 10 X buffer and 40 μL of nuclease-free water) to the column after the first wash, for a 30 minutes incubation at room temperature and final wash with the wash buffer I. The remaining of the protocol was performed as recommended by the Supplier. For conidia and germinating conidia samples, total RNA was isolated after transfer of the liquid, beads-free phase to a tube containing 1 mL of TRIzol reagent (Invitrogen). 0.25 mL of chloroform was added to each sample and the tubes were incubated for 5 minutes at room temperature and then vortexed. The tubes were centrifuged at 16000 g for 15 minutes. The aqueous phase (*approx.* 750 μL) was transferred to a clean tube and 1 volume of isopropanol was added. The samples were mixed by inverting the tubes several times. The tubes were incubated at room temperature for 10 minutes and centrifuged at 16000 g for 10 minutes. The supernatant was removed and the pellet was washed with 1 mL of 70% (v/v) ethanol and centrifuged once again at 16000 g for 10 minutes. The ethanol was discarded and the pellet was left to dry. Each pellet was resuspended in 50 μL of nuclease-free water. A clean-up protocol using the RNeasy Mini Kit (Qiagen) and on-column DNase I treatment (Thermo) was then performed on these RNA samples.

The quantification of the RNA samples was assessed by using the ND-1000 UV-visible light spectrophotometer (NanoDrop Technologies) while the Bioanalyzer 2100 with the RNA 6000 Nano LabChip kit (Agilent) was used to certify RNA integrity. Only RNA samples with 260/280 nm wavelength ratio of approximately 2 and 260/230 nm wavelength ratio greater than 2 were retained for analysis. Synthesis of cDNA was performed using the PrimeScript First Strand cDNA Synthesis Kit (Takara), following the Manufacturers' protocol. One microgram of total RNA from mycelia samples and 100 ng of total RNA from conidia and germinating conidia were used for the cDNA synthesis reaction. The cDNA was diluted 1:10 with water and stored at $-20\text{ }^{\circ}\text{C}$.

Primer design and validation

Primers were designed using Vector NTI advance v11 (Life Technologies) with melting temperature of 58–60 °C, length of 18–25 bp and GC content of 50–60%. All except R7 and R9 (see table 1) possess one to several introns in their sequence, which allowed designing the primers at the exon-exon junctions to minimise the amplification of contaminant gDNA. Amplicon sizes ranged between 70 and 200 bp. Reaction efficiency for each pair of primers was tested by the dilution series method using a mix of cDNA samples as the template. The efficiency of validated primer pairs focused around 100% (Table 1).

qPCR

The qPCR was performed in a CFX96 Real Time PCR Detection System (Bio-Rad), using 96-well white PCR plate (Thermo) sealed with ABsolute qPCR seals (Thermo). The reaction mix consisted of 7.5 µL of the DyNamo ColorFlash SYBR Green master mix (Thermo), 300 nM of each primer and 3 µL of the 1:10 diluted cDNA in a final volume of 15 µL. The PCR reaction cycle was: initial denaturation for 7 min at 95 °C, followed by 40 cycles of 10 seconds at 95 °C and 30 seconds at 60 °C. A melting curve was performed at the end of the qPCR run, increasing the temperature in a stepwise fashion by 0.5 °C every 5 seconds, from 65 °C to 95 °C. Each RT-qPCR reaction was performed in technical triplicate. Two control samples were included for each primer pair tested; the no template control (NTC) and *T. versatilis* genomic DNA. For each sample, a ValidPrime Assay (VPA), consisting of a pair of primers that bind to a non-transcribed intergenic region identified from RNA-seq data, was also included to detect and quantify the presence of contaminating gDNA (Laurell *et al.*, 2012). The primers for the VPA were; 5'ACCGAATGGCACCGAGTTGG 3' and 5'AATGGAGGAAGCGTGCCGTG 3'. As gDNA contamination rarely exceeded 1%, the RT-qPCR data were directly analysed using the CFX Manager software (Bio-Rad).

Stability analysis

The stability of putative reference genes was assessed using the geNorm VBA applet for Microsoft Excel (Vandesompele *et al.*, 2002). geNorm allows the calculation for each reference gene of the gene expression stability value M , which is the average pairwise variation of a particular gene with all other control genes, the most stable genes presenting the lowest M values. To determine the optimal number of genes that are required for an accurate normalization, the normalization factors (NF_n , based on the geometric mean of the n most stable genes) were calculated by stepwise inclusion of the most stably expressed genes. Pairwise variations ($V_{(n/n+1)}$) between NF_n and NF_{n+1} were then calculated to determine the effect of adding the $(n+1)^{th}$ gene. If the $V_{n/n+1}$ is superior to the cut-off value 0.15, the addition of the $(n+1)^{th}$ gene has a significant effect on normalization quality and should preferably be included for calculation of a reliable normalization factor.

***In-silico* analysis of RNA-seq data**

Three RNA-seq datasets from the industrial *T. versatilis* were at our disposal (unpublished data) and were prepared from: 1) growth of the mycelium on MM for 48 h (reference condition); 2) transfer of the water-rinsed mycelium to MM with ball-milled wheat straw 1% (w/v) as carbon source and sampling after 24 h; 3) direct addition of glucose at 1% final concentration to the mycelium exposed to wheat straw, and sampling after 5 h. These RNA-seq data were used for the pre-selection of stable genes (fold change (FC) equal to one) after calculating the FC as follow: RPKM (Reads Per Kilobase of exon model per Million mapped reads) value in the sample of interest / RPKM in the reference condition, for each gene. Similarly, FC for candidate reference genes were calculated from RNA-seq data publicly available at the NCBI GEO database (Barrett *et al.*, 2013; Edgar *et al.*, 2002). To identify the homologues of *T. versatilis* selected reference genes in the different fungi, a standard protein BLAST (blastp) using the amino-acid sequence from *T. versatilis* was performed against protein databases, specifying the organism. Each homologous sequence was then used for a reciprocal BLAST against the *T. versatilis* database in order to confirm the accuracy of the result. The

detailed list of locus tags for each gene in every fungus is available in the Additional file 2. For each GOI in these studies, the ratio between the expression in a condition of interest and the expression in the control condition was calculated. Collected datasets were from *Trichoderma reesei* ((Ries *et al.*, 2013), accession #GSE44648), *Aspergillus niger* ((Delmas *et al.*, 2012), #GSE33852), *Aspergillus flavus* ((Lin *et al.*, 2013; Yu *et al.*, 2011), #GSE40202 and #GSE30031), *Aspergillus fumigatus* (#GSE30579), *Aspergillus oryzae* ((Wang *et al.*, 2010a), #GSE18851), *Aspergillus nidulans* ((Coradetti *et al.*, 2013), #GSE44100), *Blumeria graminis* ((Hacquard *et al.*, 2013), #GSE43163), *Colletotrichum graminicola* ((O'Connell *et al.*, 2012), # GSE34632), *Colletotrichum higginsianum* ((O'Connell *et al.*, 2012), #GSE33683), *Fibroporia radiculosa* ((Tang *et al.*, 2013), #GSE35333), *Magnaporthe oryzae* ((Soanes *et al.*, 2012), #GSE30327), *Neurospora crassa* ((Coradetti *et al.*, 2013), #GSE44100), ((Coradetti *et al.*, 2012), #GSE35227), ((Znameroski *et al.*, 2012), #GSE36719), *Pyronema omphalodes* ((Traeger *et al.*, 2013), #GSE41631), *Rhizoctonia solani* ((Zheng *et al.*, 2013), #GSE32577), *Sordaria macrospora* ((Teichert *et al.*, 2012), #GSE33668). We also got access to unpublished data from *Rhizophagus irregularis* ((Blackwell, 2011), #SRX375378 at NCBI Short Read Archive) and *Leptosphaeria maculans* (personal communication from T. Rouxel, INRA-Bioger, Thiverval-Grignon, France).

Gene expression and Statistical analyses

Three independent cultures of *T. versatilis* were carried out to perform RNA-seq. For reference gene stability analysis by RT-qPCR, cultures of *T. versatilis* in the different conditions were performed in duplicate. qPCR assays were performed in technical triplicates. Inter-run calibrators were included in each qPCR plate. The RT-qPCR data were directly analysed using the CFX Manager software (Bio-Rad), which allows inter-run calibrations, efficiency correction, normalization with multiple reference genes and calculation of ratios with (technical) errors propagation. As advised for final calculation of FC values from biological replicates (Derveaux *et al.*, 2010), statistics (mean \pm SD) were assessed from FC values obtained from biological replicates (Fig. 4B). Other statistical analyses were conducted by using the STATGRAPHICS Centurion 16 software. This included: the ANOVA on relative FC values

presented in Fig. 3 (Three-level, nested ANOVA with 'genes' as the first level, 'culture conditions' as the second level and 'biological replicates' as the third level); the Hierarchical Ascendant Classification (HAC) that was performed according to the Ward method, using default parameters (standardization of the data and squared Euclidean distances); the graphical representation of box plots (Fig. 1 and 6)

Results & Discussion

Selection of candidate reference genes from *T. versatilis* RNA-seq datasets

In a preliminary study on the industrial *Talaromyces versatilis* strain IMI378536, RNA-seq data were generated to analyse the transcriptome of this filamentous fungus on wheat straw (unpublished data, property of ADISSEO SAS). An *in silico* screen for genes that showed no differential expression between glucose and wheat straw, was used to select about a hundred genes with fold-changes close to one, indicating stability of transcript levels under those conditions. From this pre-selection, genes were discarded because of anti-sense transcription, alternative splicing events, as well as very low expression level (RPKM below 30). The design of primers and their experimental validation by RT-qPCR were then performed on a residual list of 20 candidate genes taking care to avoid their participation in similar cellular functions to minimise the risk of co-regulation under culture conditions of interest. Finally, 12 putative reference genes were selected whose primers led to good reaction efficiency. As shown in Table 1, this selection included genes involved in intracellular signalling, vesicular trafficking, metal transport, cytoskeleton organization or protein ubiquitinylation, but quite surprisingly, it did not contain any gene implicated in the central carbon metabolism. To this list, the gene encoding β -tubulin was also included, as it is frequently used for RT-qPCR data normalization (Fernandez *et al.*, 2012; Georgakopoulos *et al.*, 2012; Lozano-Ojalvo *et al.*, 2013; McKelvey & Murphy, 2010; Raggam *et al.*, 2011).

Evaluation of the stability of candidate genes expression, in *T. versatilis* cultivated under a large set of conditions

To evaluate whether the 13 candidate genes harboured a stable transcript level and could be used as proper internal control for data normalization in RT-qPCR gene expression analysis, their transcript levels were quantified by this technique in more than 30 different conditions (Additional file 1). Growth was explored in the presence of different carbon sources (from monosaccharides to complex plant cell wall polymers), temperature, pH and salt stresses, as well as to carbon and nitrogen starvation. In addition, transcript levels of these genes was monitored during conidial germination, as this developmental process is a particularly interesting aspect in fungal biology (Lamarre *et al.*, 2008; Leeuwen *et al.*, 2013; Novodvorska *et al.*, 2013). The raw Cq values of the 13 genes were therefore collected under 33 conditions and compiled in the box plot (Fig. 1). Most of these genes showed a compact distribution of Cq values, with less than 2 Cq between the 1st and 3rd quartiles, indicating relatively low variation of the transcript level among the different conditions (*i.e.* less than 4-fold differential expression for the middle fifty). Some of them, R3 (*ADA*), R11 (*AP-2B*) as well as the *β-tub* gene displayed slightly higher dispersion of their Cq values. The genes R3 and *β-tub* also exhibited weaker and stronger transcript levels, respectively, with *approx.* 100-fold differential average transcript level between each other. Besides these 2 candidates, quite similar average expression levels for the remaining 11 genes, with raw Cq values around 25, were observed. This average expression level was acceptable for robust RT-qPCR assays and normalization, based on the validated reaction efficiencies and the possibility to amplify target cDNA over several logs of concentration. However, this was contrary to the notion that the transcript level of the ideal reference gene must be close to the average transcript level of the GOIs. That situation cannot occur when the GOIs present very different average transcript levels, or when a single GOI presents either potent repression or strong induction in the same study. As an example, the expression of *abf-B2* encoding a GH54 α-L-arabinofuranosidase (Guais *et al.*, 2010), showed more than 25-fold relative

Appendix

change between the 1st and 3rd quartiles. The huge whiskers of *abf-B*'s box (Fig. 1) reflected more than four log differential expression, and at least a 5-log differential expression of *abf-B2* between the two extreme conditions was observed.

The raw Cq values were transformed to quantities with efficiency correction and then analysed with the geNorm algorithm to rank the 13 candidate genes according to their *M* value and to ascertain their expression stability over a specified set of conditions (Fig. 2A). When considering the whole set of conditions that were investigated (“all conditions”), the *M* value of these 13 genes (numbered 1 to 13 on the X axis in Fig. 2A) was below the recommended threshold of 1.5, even for the *β-tub* gene, but also R3 and R11, which ranked amongst the least stable genes in agreement with the behaviour that was reported in Fig. 1. To reinforce this result, RefFinder (<http://www.leonxie.com/referencegene.php>), a web-based tool that integrates the currently available major computational programs (geNorm (Vandesompele *et al.*, 2002), Normfinder (Andersen *et al.*, 2004), BestKeeper (Pfaffl *et al.*, 2004) and the comparative Δ Ct method (Silver *et al.*, 2006)) was also used, and it led to almost identical classifications (data not shown). Still using geNorm, the optimal number of reference genes required for accurate normalization in the “all conditions” set (Fig. 2B) was evaluated. Vandesompele and coworkers (Vandesompele *et al.*, 2002) recommended a cut-off value at 0.15 for the pairwise variation value ($V_{n/(n+1)}$), below which the inclusion of an additional gene does not result in a significant improvement of the normalization. According to this criterion, the $V_{2/3}$ and $V_{3/4}$ values indicated that three genes, *i.e.* R2 (*ubcB*), R10 (*sac7*) and R6 (*psm1*), were sufficient for accurate normalization of transcript levels in any of the samples examined.

Evaluating gene expression stability in subsets of conditions

The stability of transcript levels was similarly analysed in subsets of selected conditions, *i.e.* samples from mycelium grown in different carbon sources (‘C sources’ subset), samples from mycelium exposed to different stress conditions (‘Stress’ subset) and samples harvested during

conidial germination ('Germination' subset) (Fig. 2A). While R1, R11 and R3 were amongst the least stable genes when analysing the whole set of conditions, they classified amongst the 3 best genes in the 'C sources', 'Stress' and 'Germination' subsets, respectively. This reorganisation of the ranking, when conditions changed, could be explained by the uniformly stable transcript levels from these genes, particularly in the 'C sources' and 'Stress' subsets, which led to low and stable M values for 11 and 9 genes amongst the 13 candidates, respectively. This remarkable stability for most of the genes was also supported by the pairwise variation values (Fig. 2B), which indicated that only two genes could ultimately be used for robust normalization in these 'C-sources' (R6 and R2), 'Stress' (R7 and R11) and 'Germination' (R10 and R2) subsets. In the first two subsets, the stepwise inclusion of reference genes led to continuous decrease of the V value, until the inclusion of the least stable genes reversed the tendency. These values were nevertheless always below the 0.15 cut-off value, confirming the extreme stability of transcript levels from these genes.

In the context of conidial germination, the identification of reliable reference genes was first challenged by the difficulty of producing good quality RNA samples. The influence of RNA quality on gene expression stability was recently reviewed (Die & Román, 2012; Vermeulen *et al.*, 2011), highlighting that the process of normalization does not completely resolve the bias of using compromised RNA quality on the final results. In our hands, only the use of the TriZol reagent secured the mRNA quality standard required for reliable RT-qPCR analysis. This technical prerequisite being fulfilled, the analysis of C_q values in this 'Germination' subset showed that the M values increased more rapidly than for 'C-sources' and 'Stress' subsets, indicating higher expression variability of the genes. This was further illustrated by a hierarchical ANOVA of the relative transcript level data (Fig. 3), where it was observed that 70 to 80% of the variation for the 'C-sources' and 'Stress' subsets took place at the level of the biological replicates (Fig. 3A & B), supporting the extremely low variation between genes as well as the low influence of conditions on the transcript levels. In contrast, the variation observed between genes strongly increased in the 'Germination'

subset, to reach about 50% of total variation (Fig. 3C), which was particularly emphasized with genes such as R6 and R11 that exhibited a strong bias (higher expression and activation during germination). The genes R10, R2 and R3, which were classified by geNorm as the best reference genes in this specific subset, were used for normalization (see below, $NF_{(R10, R2, R3)}$) and confirmed that R6 and R11 were induced respectively by 6 and 12-fold, 6 hours after the beginning of the germination process (data not shown).

Robust normalization using the geometric mean of a minimal number of these candidate reference genes

A safety solution to certify a good normalization factor and minimise the risk of bias if using a single gene, is to normalize using the geometric mean of multiple reference genes, as it was proposed previously (Vandesompele *et al.*, 2002). The overall good stability of the candidate reference genes prompted a further challenge to their suitability for normalization, by evaluating the normalization bias that might result from the use of sub-optimal combinations of these putative candidates. This normalization bias, *i.e.* under- or over-estimation of the normalised expression value of GOIs, was calculated as the ratio between a given normalization factor (NF) and the ideal one ($NF_{(R10, R2, R6)}$). Fig. 4A reported results from different combinations of three candidates, in this case from less and less well ranked genes ($NF_{(R12, R7, R8)}$, $NF_{(R4, R9, R1)}$ and $NF_{(R5, R11, R3)}$, respectively). In most cases, these NFs led to a minimal normalization bias and only sporadic samples showed 2.0–3.0-fold down-estimation, especially when using the three least stable candidate genes from this study (conditions #12, 17, 20 and 21 from the $NF_{(R5, R11, R3)}$ series, blue zone). Although not perfect, this latter NF was nevertheless far better than the use of R3 alone, which led to 3.0–8.0-fold under-estimation in about half of the samples tested in our study (Fig. 4A, lower red zone). Similarly to R3, the β -*tub* gene that ranked amongst the least stable genes (Fig. 2A) strongly altered the quality of the normalization with a strong over-estimation in most of the conditions investigated (Fig. 4A, blue and red zones). This significant reduction of the normalization bias by averaging R3 in $NF_{(R5, R11, R3)}$ was

therefore a clear illustration of the prime importance to normalise by the geometric averaging of multiple genes (Vandesompele *et al.*, 2002). Ideally, the optimal number of genes relies on the stepwise inclusion of additional reference genes until the time when this supplementary gene may not improve, nor deteriorate the normalization factor (Vandesompele *et al.*, 2002). If the genes are not highly regulated and, more importantly, not co-regulated, it was confirmed that any combination of 3 of the genes could lead to very similar NF values and minimal normalization bias.

The transcript level of two relevant genes of *T. versatilis*, namely *abf-B2* encoding a GH54 arabinofuranosidase (Guais *et al.*, 2010) and *xynC* encoding a GH11 xylanase (Lafond *et al.*, 2014), under three different culture conditions (Fig. 4B) was explored to further illustrate the normalization bias. Using the ideal normalization factor ($NF_{(R2, R10, R6)}$), it was found that the expression of *abf-B2* increased by 9.7 and 33.3 fold upon transfer from glucose to thio-gentiobiose and Arbocel, respectively, whereas the transfer to C-starved medium did not cause any significant change of transcript level. A completely different regulatory pattern was obtained using a single reference gene for normalization. The use of R3 led to under-estimation of *abf-B2* transcript level upon transfer to thio-gentiobiose, while, and even worst, it indicated that this gene was clearly repressed upon C-starvation. Only in a few circumstances (*e.g.* Arbocel sample), the use of single genes for normalization (both R3 and β -*tub*) led to similar fold change values. Likewise, transcript level analysis of *xynC* showed that a single reference gene could lead to erroneous conclusions. After normalization by R3, the absence of differential expression upon transfer to thio-gentiobiose could be concluded and slight repression upon C-starvation, while the conclusion might have been strong repression and similar overexpression of *xynC* in these two culture conditions when using the β -*tub* gene as the internal control. Taken together, these few examples clearly support the need for identification, validation and use of multiple reference genes to avoid incorrect biological interpretation of gene regulation, particularly if the biological significance of subtle differences in fold-changes values is to be considered.

Expression data collection from phylogenetically distant filamentous fungi

To demonstrate the suitability of some of the 12 putative reference genes for RT-qPCR analysis in fungi, RNA-seq based transcriptomic datasets from 18 phylogenetically distant filamentous fungi were interrogated, taking benefits of web resources such as the GEO portal (Barrett *et al.*, 2013; Edgar *et al.*, 2002). These datasets covered model fungi (Coradetti *et al.*, 2012, 2013; Teichert *et al.*, 2012; Traeger *et al.*, 2013; Znameroski *et al.*, 2012), biotechnologically important organisms (Delmas *et al.*, 2012; Ries *et al.*, 2013; Wang *et al.*, 2010a), agronomical relevant fungi such as symbiotic organisms or plant pathogens (Hacquard *et al.*, 2013; O'Connell *et al.*, 2012; Soanes *et al.*, 2012; Tang *et al.*, 2013; Zheng *et al.*, 2013), as well as human pathogens (Lin *et al.*, 2013; Yu *et al.*, 2011). This collection corresponded to 94 independent datasets, most of them in triplicates and harvested from a broad variety of experimental conditions, *e.g.* exposure to stress, nutritional source utilisation, fungi-host interactions or development stages (see Additional file 3 for further details). To further strengthen this analysis, genes were included that have been evaluated for their use in *Aspergillus niger* (Bohle *et al.*, 2007) and *Trichoderma reesei* (Steiger *et al.*, 2010). The 12 genes selected here are henceforth referred to as the 'R series' and the additional reference genes were termed 'C series' and contained the actin (C1, *act*), aminopeptidase C (C2, *apsC*), cytochrome C oxidase subunit V (C3, *coxV*), glyceraldehyde-3-phosphate dehydrogenase (C4, *gapdh*), glucokinase (C5, *glkA*), glucose-6-phosphate dehydrogenase (C6, *g6pdh*), isocitrate dehydrogenase precursor (C7, *icdA*), phosphofructokinase (C8, *pfkA*), phosphoglucose isomerase (C9, *pgiA*), a secretion associated GTP-binding protein (C10, *sarA*), and the translation elongation factor α 1 (C11, *ted1a*). The β -*tub* gene (C12), which was not part of *A. niger* and *T. reesei* studies, was also included in this study as it has been evaluated amongst other putative reference genes in similar studies ((Fang & Bidochka, 2006; Hacquard *et al.*, 2011; Raffaello & Asiegbu, 2013; Vieira *et al.*, 2011; Zampieri *et al.*, 2014; Zhou *et al.*, 2012) and Additional file 4). Finally, 4 further genes were added that were homologous to *S. cerevisiae* *ALG9* (Sc1), *TAF10* (Sc2), *TFC1* (Sc3) and *UBC6* (Sc4), which have been previously validated as good reference genes in this yeast (Teste *et al.*, 2009). For every gene, in each specific study, fold

changes (FC) were calculated as the ratio between the expression in a condition of interest and the expression in the control condition that was designed in this specific study.

Changes in transcript levels in the different conditions for the different fungi, were used to generate a heat map (Fig. 5). None of the genes from the 'C series' was stable in the two *T. versatilis* conditions and explains why they were not pre-selected in this study. Also, in RNA seq data collected from *R. irregularis*, *P. omphalodes*, *R. solani*, *F. radiculosa*, *M. oryzae* and *A. nidulans*, almost all the candidate genes and those from the 'C series' exhibited good expression stability, as indicated by the generalized greenish colour. Half of those genes had FC values lower than 1.2 (data not shown) and FC values almost never higher than 2. During the early stages of plant infection by *B. graminis*, extensive transcriptomic changes were observed for most of the genes. A similar situation was observed during different development stages of *S. macrospora*, where most of the reference genes were either down- or up-regulated. These two datasets strengthened the conclusion that even the best reference genes will never escape sporadic differential expression, so that validation of their stability is highly recommended prior to their use for normalization in each new project. When focused on the different *Aspergillus* and *Neurospora* species, visual inspection of the heat map suggested that genes from the 'R series' were slightly more stable. This was confirmed when pooling and analysing together all FC values from the 'R' and 'C' series, respectively, as the median and interquartile of these two gene subsets indicated a clear tendency to down regulation of the 'C series' (results not shown). A striking feature within this 'R series' was nevertheless the R5 (*Cu-ATPase*) gene that was strongly regulated under a few conditions, particularly in *N. crassa* exposed to Avicel or carbon starvation. This latter observation was however specific to *N. crassa*, as this gene was perfectly stable in *T. versatilis* mycelium similarly exposed to this carbon source or to C starvation (not shown). Finally, the fungal genes homologous to yeast Sc3 (*TFC1*) and Sc4 (*UBC6*) exhibited relatively stable transcript levels in most of the filamentous fungi and conditions of

interest, with the notable exception of *B. graminis* during plant infection as already mentioned above. In deep contrast, Sc1 (*ALG9*) exhibited much higher fluctuations of FC values.

Global analysis of reference genes stability in filamentous fungi

To find out the most relevant reference genes among the whole 'R' and 'C' series, the 94 FC values were pooled for each gene independently (Fig. 6). When looking at the median and interquartile, which are robust statistical parameters especially for rather small samples that are not normally distributed, the most promising genes should present a median close to zero and a compact interquartile, indicating no differential expression and low variation, respectively. Remarkably, R2 (*ubcB*), which was identified by geNorm as the best reference gene for *T. versatilis*, exhibited such requirements with a median close to zero and FC values that did not exceed 1.3 (repression or induction) in half of the conditions collected in this study. In contrast, the R5 gene (*Cu-ATPase*) exhibited a very strong bias towards overexpression and a much higher variation in FC values, even when removing outliers that mostly fitted with *N. crassa* samples exposed to cellulose.

To search for groups of genes presenting similar behaviour, a cluster analysis (HAC) was conducted, using the median and the interquartile as variables (Additional file 5). Moderate partitioning led to the identification of 5 classes, highlighted in the interquartile *versus* median scatter plot (Fig. 7). The best class (red), with its centroid having a median at zero and the lowest variation, contained three of the new candidates (R2 (*ubcB*) and R10 (*sac7*), previously designated as the best reference genes for *T. versatilis*, and R4 (*fis1*)), C10 (*sarA*) that encodes a secretion-associated GTP-binding protein that was already identified as a good reference gene (Bohle *et al.*, 2007; Steiger *et al.*, 2010), and two genes homologous to *S. cerevisiae* *TFC1* (Sc3) and *UBC6* (Sc4). In contrast, the orange class, which had the largest interquartile of the study indicative of poor stability and higher probability of differential expression, included C8 (*pfk*), C4 (*gapdh*) and C2 (*apsC*) genes. In between, the green category contained several classic reference genes such as a C1 (*act*) and C5 (*glk*), which presented quite centred medians but showed the highest variability within this class.

Other frequently used reference genes such as C12 (*β-tub*), C11 (*tef1*) and C6 (*g6pdh*) were categorised in a less promising group that presented a reasonable level of variability but a tendency towards down-regulation (class 4, blue). Finally, R5 (*Cu-ATPase*) and R3 (*ADA*), which classified amongst the least stable genes in *T. versatilis* samples, confirmed their low stability in other filamentous fungi and a clear bias to overexpression (class 5, violet).

Even though genes in the red class could be considered as a very promising set of reference genes for normalization purposes in fungi, they too may present unexpected regulatory changes in specific contexts. A very interesting case-study was indeed found in this work with two of these most promising reference genes, *i.e.* R4 (*fis1*) and Sc4 (*UBC6*), which were strongly down-regulated in *C. higginsianum* during infection phases of *Arabidopsis thaliana*. Moreover, although a reference gene is certified in an organism of interest, it does not preclude extremely different regulatory patterns in phylogenetically distant organisms exposed to strictly similar conditions. It was found for example that R5 (*Cu-ATPase*) was stable in *T. versatilis* exposed to Avicel or C starvation (data not shown), while it was strongly activated by these environmental conditions in *N. crassa*. Such examples clearly emphasise that validation experiments are mandatory to avoid the drawbacks of using inappropriate reference genes (Dhedda *et al.*, 2005; Faccioli *et al.*, 2007). The second important point that should be stressed is the possible co-regulation of selected reference genes. While apparently linked to GTP, R10 (*sac7*) and C10 (*sarA*) seem to be implicated in independent functions, *i.e.* signal transduction and secretion, respectively. Unfortunately, this is not the case for R2 (*ubcB*, Ubiquitin-protein ligase activity) and Sc4 (*UBC6*, ER-associated protein catabolic process), which belong to a similar functional category and hence may show undesirable co-regulation. Therefore, the use of both of them as reference genes should be discouraged, although both could be evaluated.

It is unlikely that expression of most of these promising reference genes will be found to be unstable simultaneously in future projects. If this was the case, it will require identifying and validating new genes. Transcriptomic data obtained from distant organisms studied under

comparable conditions, or from the organism of interest cultivated in conditions as diverse as possible, could be collected. This strategy turned out to be successful in our hands for pre-selection of appropriate reference genes, even from a very limited set of transcriptomic data as starting material. The automated identification of suitable reference genes by the use of tools such as RefGenes (Hruz *et al.*, 2011) might be useful, taking care to focus on functionally unrelated candidates, provided that transcriptomic data that are targeted by this tool are generalised to all published datasets, including those that have been produced from filamentous fungi.

Conclusions

Starting with a restricted set of 3 RNA-seq based transcriptomic datasets from *T. versatilis* cultivated under extremely different growth conditions, a list of 12 stable genes that belong to different functional classes was selected. The stability of transcript levels in *T. versatilis* in more than thirty conditions of interest such as varied nutritional sources, stress exposure or time course analysis of conidial germination was examined. Three genes, R2 (*ubcB*), R10 (*sac7*) and R6 (*psm1*), were certified as the best reference genes for accurate normalization of expression levels, but the overall good stability of the 12 candidates led to the conclusion that any combination of 3 of them could lead to minimal normalization bias, even with the least stable genes on the list.

Beyond their robustness in *T. versatilis*, the suitability of these reference genes for RT-qPCR analysis within the filamentous fungal kingdom was assessed, by collecting RNA-seq based transcriptomic data from 18 phylogenetically distant fungal species. The representative set of organisms and experimental conditions that was collected, confirmed that most of the classic “housekeeping” genes such as *g6pdh*, *β -tub* and *act* did not appear as the most stable genes, even if the latter has been classified among the best reference genes in few specific studies (Bohle *et al.*, 2007; Raffaello & Asiegbu, 2013). Other genes involved in central metabolism, *e.g.* *pfk*, *gapdh*, *aspC* or *glk*, also showed the highest occurrence of significant down- or up-regulation, which, together

with their probable risk of co-regulation, strongly discouraged their further use as reference gene for RT-qPCR gene expression analysis.

The most promising group of six reference genes contained *ubcB* (ubiquitin carrier protein), *sac7* (Rho GTPase activator), *fis1* (mitochondrial membrane fission protein), *sarA* (secretion associated GTP-binding protein), and two genes homologous to *S. cerevisiae* *TFC1* and *UBC6* (proteins involved in transcription initiation on Pol III promoters and ER-associated protein catabolic process, respectively). Four among these six genes – *sac7*, *fis1*, *sarA* and *UBC6* – did present a non-normal distribution with rare cases of strong differential expression in these RNA-seq conditions. As already claimed, there is no universal gene that exhibits stable expression levels in any sample and/or organism of interest (Vandesompele *et al.*, 2002) and it seems utopic to think escaping from sporadic differential expression. The need for systematic validation of the stability of transcript levels from these reference genes in forthcoming studies is therefore warranted.

Ethics

Material and experiments carried out in the frame of this article did not require any ethics approval.

Competing interests:

The authors declare that they have no competing interests.

Authors' contributions:

AL, JMF and JLP conceived the study. AL carried out the cultures, samples treatment and Rt-qPCR analysis. AL and JLP analysed the results. AL, JMF and JLP wrote the manuscript and all authors approved the final version.

Acknowledgements:

This work was supported in part by the CINABio-Adisseo (Adisseo SAS, Antony, France) to JMF & JLP and by CIFRE, grant number: 1558/2010 to AL. We are grateful to Christophe Roux (LRSV, Toulouse, France) and Thierry Rouxel (INRA-Bioger, Thiverval-Grignon, France) for providing access to their unpublished RNA-seq data from *Rhizophagus irregularis* and *Leptosphaeria maculans*, respectively, Jolanda Van Munster (University of Nottingham, UK) for providing RNA-seq data analysis from *T. reesei* and *A. niger*, and David Archer, Steve Pullan, Stéphane Delmas and Martin Blyth (University of Nottingham, UK) for contribution in the production of *T. versatilis* RNA-seq results (unpublished results). We also thank David Archer (University of Nottingham, UK) for scientific support during this work and proofreading of the manuscript.

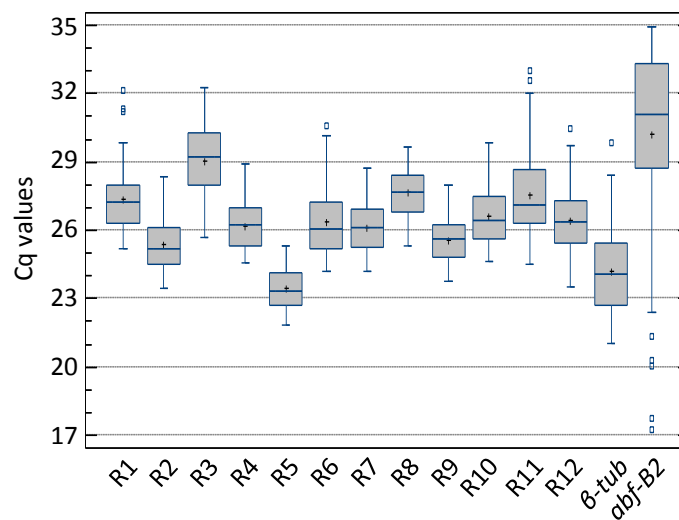


Figure 1: Distribution of the raw Cq values. For each gene, the box-plot gathered all the 66 raw Cq values obtained from the 33 duplicated culture conditions of *T. versatilis*. The lower and upper boundaries of the box (interquartile) represent the 25th and the 75th percentile, respectively. The line within the box corresponds to the median and the cross to the mean of the distribution, while the whiskers indicate the highest and the lowest Cq values, with the exception of the outliers that are represented by the squares outside the whiskers.

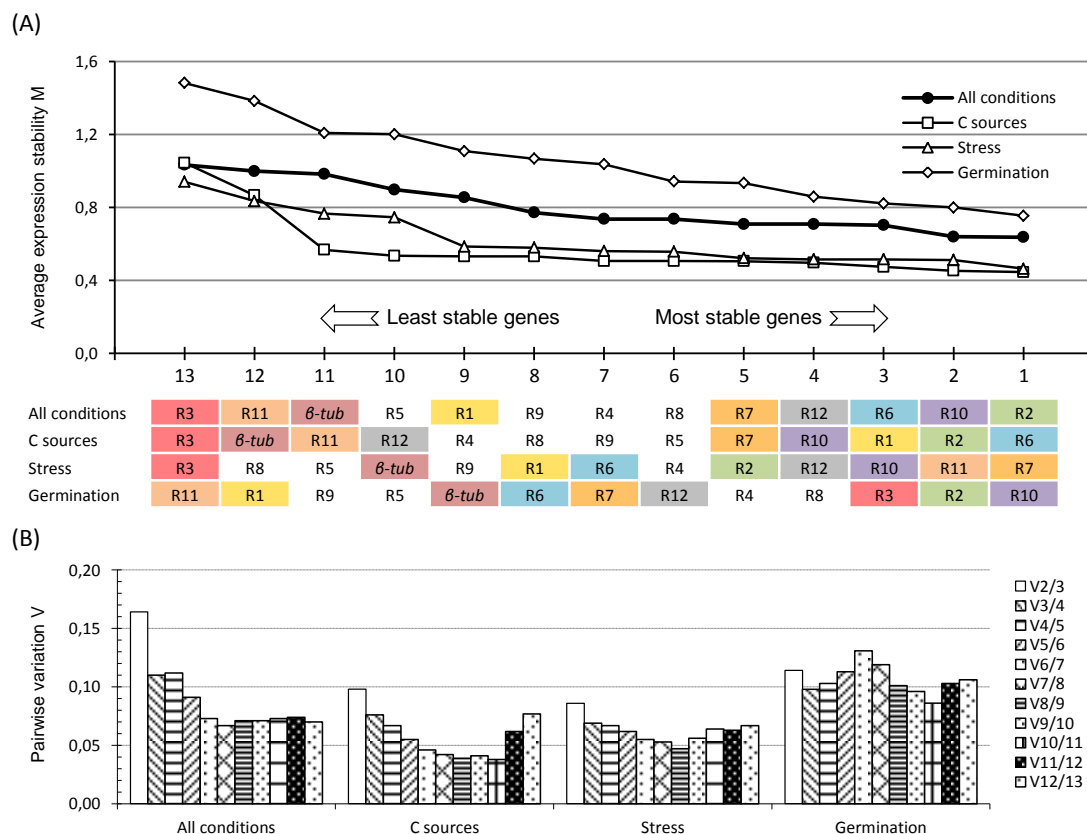


Figure 2: geNorm–based ranking of the putative reference genes. (A) Genes were ranked from the least stable (on the left) to the most stable (on the right) according to their M value (Y axis). This classification was independently performed by using different sets of conditions: the ‘All conditions’ included the whole set of culture conditions studied by RT-qPCR in this work; the ‘C sources’ subset gathered the 18 samples obtained from growth on different sugars; the ‘Stress’ subset corresponded to 6 samples harvested during stress exposition; the ‘Germination’ subset included the 6 germination time points. For each set of conditions, the result of the classification was given below the X-axis (arbitrary colours attributed to each gene for the sake of clarity). **(B)** Result of the pairwise variation analysis between NF_n and NF_{n+1} to determine the optimal number of genes for reliable normalization. Values below the 0.15 threshold mean that n genes might be sufficient.

Appendix

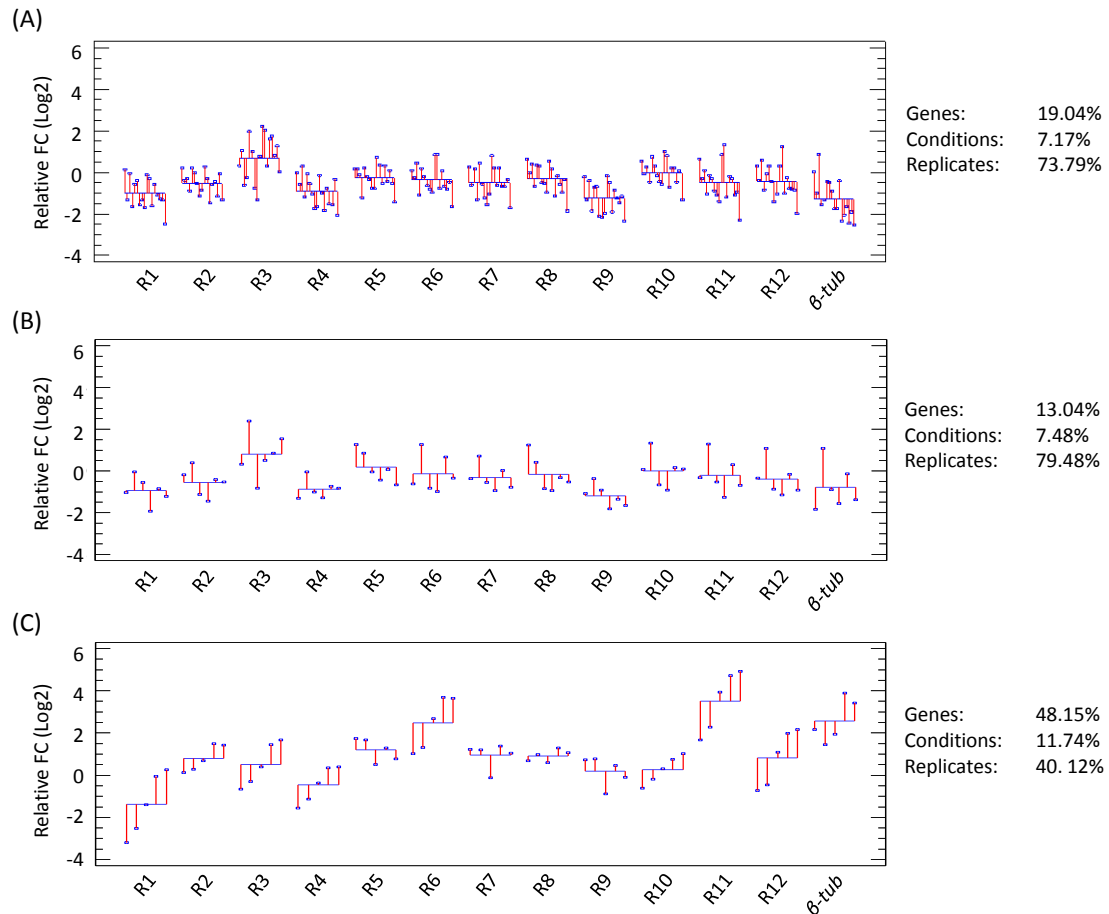


Figure 3: Hierarchical ANOVA of the putative reference genes. Three-level, nested ANOVA with ‘genes’ as the first level, ‘culture conditions’ as the second level and ‘biological replicates’ as the third level. As in Fig. 2, this ANOVA was carried out using different sets of conditions: ‘C sources’ subset **(A)**; ‘Stress’ subset **(B)**; ‘Germination’ subset **(C)**. **Left graph:** relative expression values ($\text{Log}_{(\text{base } 2)}$) as a function of the different conditions for the different genes, taking as the control conditions the glucose sample (A & B subsets) and the T0 time-point for spore germination (C subset). Two values were used for each condition (*i.e.* duplicated experiment). **Right panel:** partitioning of the variance into the three levels (in %).

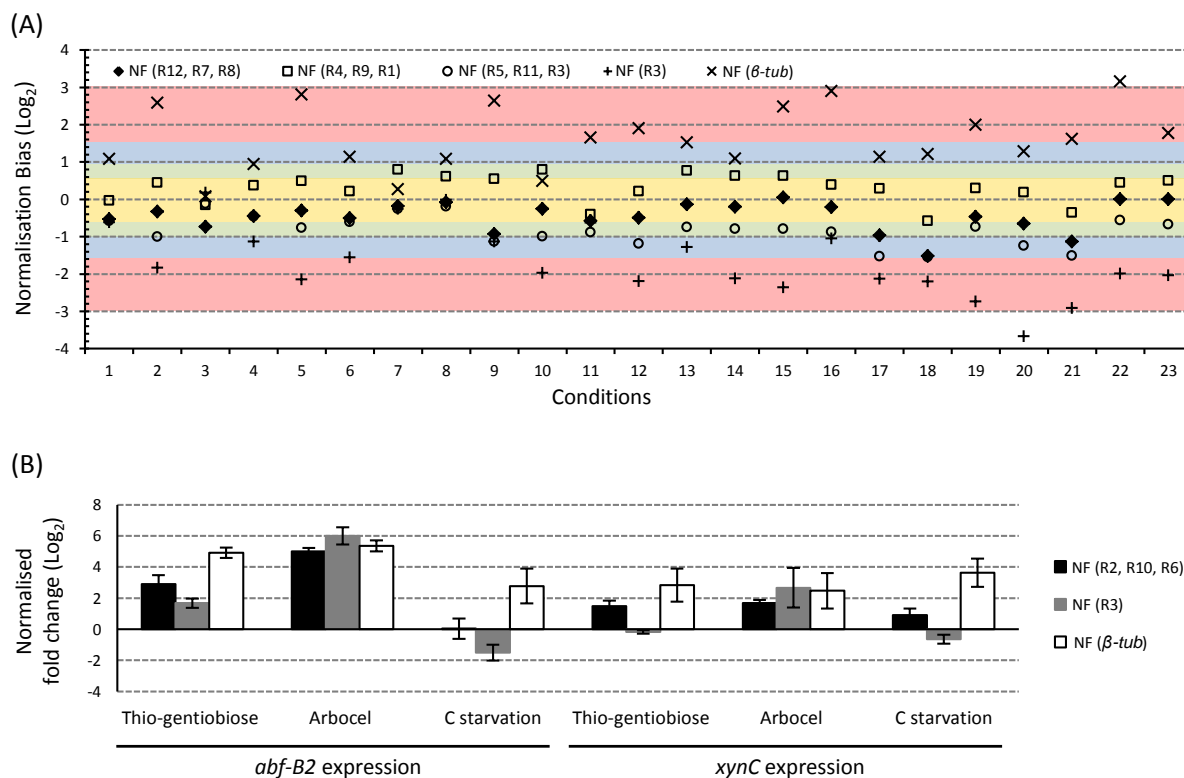


Figure 4: Normalization bias analysis according to reference genes selection. For each condition of interest (X axis, e.g. #22 was the thio-gentiobiose sample), we calculated a normalization bias (*i.e.* under- or over-estimation of the normalised expression value of GOIs) as the ratio between the theoretically best NF ($\text{NF}_{(R2, R10, R6)}$ as determined from geNorm classification by using the entire set of conditions) and NF calculated from other combinations of reference genes. $\text{Log}_{(\text{base } 2)}$ of the normalization bias is represented on the Y axis. Yellow zone: less than 1.5 fold bias; Green zone: 1.5–2 fold bias; Blue zone: 2–3 fold bias; Red zone: 3–8 fold bias. **(A)** Comparison of $\text{NF}_{(R2, R10, R6)}$ to NF calculated from less and less stable genes, as well as from single genes such as R3 and β -*tub*. **(B)** Example of quantitative expression results of *abf-B2* and *xynC* under different culture conditions (Thio-gentiobiose, Arbocel and C starvation samples) using 3 different NFs: $\text{NF}_{(R2, R10, R6)}$, $\text{NF}_{(R3)}$ and $\text{NF}_{(\beta\text{-Tub})}$. $\text{Log}_{(\text{base } 2)}$ of FC values on the Y axis (mean \pm SD, n=2 in this experiment).

Appendix

Fungi	New candidate reference genes												Classic reference genes												S. cerevisiae			
	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Sc 1	Sc 2	Sc 3	Sc 4
<i>T. versatilis</i>	-0.08	0.14	0.02	0.00	-0.06	0.01	-0.01	0.01	0.06	0.19	0.06	0.09	-0.51	3.47	-2.29	-1.59	-1.02	-1.70	0.56	-1.52	-0.04	0.49	-2.73	-0.23	-0.42	1.07	-0.19	-0.05
	0.08	0.01	-0.05	0.07	0.07	0.16	0.04	0.18	0.23	0.07	0.03	0.25	0.67	0.70	-0.10	1.64	0.88	-0.58	-0.19	0.10	1.00	0.34	-0.01	1.31	0.25	0.15	-0.07	-0.77
<i>A. niger</i>	-0.58	-0.55	-0.12	-0.09	-0.79	-1.03	-0.54	-0.31	-0.15	0.19	-0.95	-0.25	-1.36	-3.73	2.41	-2.22	-0.06	-1.15	-1.41	-2.88	-0.84	-0.42	-0.35		-1.19	-0.31	1.84	-0.34
	0.95	-1.05	-0.80	0.42	-0.08	-0.83	-1.32	-0.75	2.53	0.87	-0.48	-0.20	-0.70	-2.74	0.46	0.91	0.94	-0.39	0.74	0.31	0.27	0.64	0.43	-0.93	0.07	0.69	0.40	
<i>A. flavus</i>	0.08	-0.68	0.20	-0.32	-0.33	0.25	-0.26	-0.15	-0.19	-0.08	0.43	0.04	0.03	0.18	-0.78	0.11	0.14	0.01	-0.60	0.57	-0.29	-0.05	0.82	-0.32	-0.72	-0.36	0.78	0.40
	0.43	0.10	0.28	-0.33	-0.57	0.22	0.64	0.10	0.47	-0.14	0.14	-0.19	0.15	-0.04	-0.35	0.22	0.19	0.17	0.02	0.76	0.46	-0.09	0.36	-0.13	-0.30	-0.05	-0.03	-0.03
<i>A. flavus</i>	1.74	0.82	-1.47	-0.60	-2.52	-0.16	1.63	-0.11	-0.38	1.26	0.54	-0.37	0.40	-0.42	0.20	0.95	0.60	0.84	-1.79	2.04	0.98	0.97	-0.27	0.00	-0.09	1.08	1.18	0.75
	1.43	0.56	-1.46	-0.62	-0.49	-0.46	-1.74	2.20	-1.51	-1.49	-0.67	-0.17	-2.40	-0.89	-0.06	-0.47	-1.78	-0.94	-0.45	-0.74	-0.43	-0.44	-0.67	-3.28	-1.13	-1.33	0.09	
<i>A. fumigatus</i>	0.34	1.66	-0.52	-0.05	1.26	-0.47	-0.27	-0.40	0.74	-0.15	-0.39	-0.22	-0.45	-1.07	-0.40	-0.35	-0.38	-0.61	-0.46	-0.29	-0.62	-0.22	-0.62	-0.65	-0.74	-0.45	-0.16	-0.18
	0.74	0.38	-1.01	-0.69	-0.67	-0.44	-0.11	1.23	-0.08	-0.44	-0.26	-3.83	-3.14	-1.03	-0.09	0.02	-1.53	-0.76	-0.23	-1.24	-0.42	-0.61	-1.03	-0.77	-0.14	0.18	-0.38	
<i>A. oryzae</i>	0.27	1.59	0.64	0.03	-0.32	-0.02	0.53	1.32	0.01	0.13	-0.11	-0.66	-0.35	-0.07	-1.04	0.16	-1.32	-0.39	-1.85	-1.08	0.67	-0.56	-0.87	0.13	0.04	0.22		0.68
	0.18	0.84	-0.50	0.90	1.18	-0.03	-0.09	1.18	1.79	0.11	-0.66	2.09	-0.47	1.00	-1.60	-0.59	-1.93	0.47	-1.16	-1.54	-0.11	0.74	-0.76	-1.57	-0.29	0.20	-0.02	-0.06
	-0.32	1.06	0.34	0.60	1.12	-0.58	-0.26	2.02	3.21	0.32	-1.17	0.99	-1.52	0.86	-2.59	-0.76	-2.97	0.16	-1.70	-1.83	-0.38	0.17	-2.44	-2.23	-1.58	-0.25		-0.66
<i>A. nidulans</i>	-0.75	-0.49	0.22	-0.78	-0.37	-0.17	-0.98	-0.26	0.52	0.08	-0.47	-0.86	-0.05	0.48	-0.17	0.46	0.85	0.38	1.03	-0.23	0.25	-0.76	-0.11	-0.24	-0.93	-0.24	0.03	-0.13
	0.55	0.09	0.44	0.41	0.22	0.16	0.63	0.39	-0.23	-0.01	0.16	0.63	-1.07	-0.25	0.07	-0.27	-1.15	-0.12	-0.17	0.21	-0.14	0.39	-0.11	-0.93	0.62	0.20	-0.23	0.17
	-0.55	-0.02	0.90	0.01	0.97	-1.99	-1.32	-1.22	-1.32	-1.26	-1.04	0.53	-1.02	-1.19	-1.49	-1.10	-0.89	1.82	0.34	-1.44	-1.50	-1.48	-0.24	-1.22	-0.52	-1.48	0.08	-0.82
	-0.98	-0.80	0.72	-0.29	0.82	-2.30	-0.98	-1.42	-0.68	-0.82	-1.17	-1.09	-1.83	-0.64	-1.51	-2.04	-1.17	-0.26	0.11	-2.05	-1.49	-1.96	-0.52	-1.10	-0.96	-1.54	-0.94	-1.55
	-0.39	-1.05	0.17	-0.69	-0.27	-2.19	-0.93	-1.17	-0.85	-1.21	-1.64	-2.07	-1.66	-0.99	-1.17	-2.36	-0.91	-1.06	-0.36	-2.42	-1.00	-1.58	-0.70	-0.72	-1.26	-1.28	-0.84	-1.52
	0.13	-0.54	1.72	0.94	1.29	-1.17	-0.98	-0.77	-1.54	-0.57	-1.08	0.83	-0.59	-1.22	-1.23	-1.01	-0.56	2.34	0.63	-1.18	-1.20	-1.03	0.27	-1.79	-1.57	-1.52	-0.97	-1.80
	-0.15	-0.40	1.72	0.16	1.11	-1.48	-0.89	-0.62	-0.35	-0.39	-1.51	-0.98	-1.28	-0.77	-1.23	-1.95	-0.80	-0.05	0.50	-2.18	-1.05	-1.43	-0.04	-0.19	-1.86	-1.17	-0.76	-1.35
	-1.14	-1.75	0.61	-0.62	-0.98	-2.39	-1.56	-1.83	-1.60	-1.62	-2.46	-2.31	-2.11	-1.79	-1.94	3.21	-1.56	-1.50	-0.75	3.66	-1.73	2.35	-1.14	-1.96	-2.47	-2.02	-1.91	-2.39
<i>B. graminis</i>	-1.61	-1.53	0.49	-0.89	-0.23	-1.81	-1.45	-1.86	-2.51	-1.76	-1.72	-0.29	-1.44	-1.75	-2.38	-2.33	-1.76	0.51	-1.05	-2.58	-1.96	-2.25	-0.81	-0.63	-0.85	-1.03	0.03	-0.51
	-1.57	-1.50	0.34	-1.21	-0.01	-1.60	-2.22	-1.35	-1.42	-1.10	-1.40	-1.10	-1.81	-1.10	-1.78	-2.42	-1.36	-0.85	-1.03	-2.22	-1.83	-2.56	-0.77	-1.08	-2.18	-1.07	-0.64	-1.10
	-1.95	-2.34	-0.16	-1.65	-1.20	3.09	-2.40	-2.46	-2.11	-2.21	2.70	3.18	-2.63	-2.15	-2.98	3.44	-1.76	-2.72	-1.90	-3.68	-2.13	-2.98	-2.04	-1.45	-1.25	-2.41	-1.45	-2.82
	-1.03	-0.45	1.47	-0.05	0.50	-0.71	-1.22	-1.47	-1.37	-0.87	-0.88	0.44	-1.13	-0.42	-1.56	-1.69	-0.85	1.41	-0.03	-1.63	-1.42	-1.97	0.07	-0.67	-0.28	-0.46	-0.02	-0.90
	-1.20	-1.14	1.29	-0.86	0.09	-1.11	-1.29	-0.92	-1.69	-1.14	-1.26	-0.84	-1.74	-1.24	-1.54	-2.13	-0.85	-0.17	-0.68	-1.88	-1.19	-1.80	-0.37	-0.96	-1.12	-0.80	-0.97	-1.63
	-0.29	-0.69	2.31	-0.54	-0.08	-1.08	-0.81	-0.31	-0.52	-0.29	-0.51	-0.76	-0.85	-0.32	-0.79	-1.67	-0.04	-0.37	0.32	-1.81	-0.01	-1.18	0.34	0.23	-0.71	-0.56	0.00	-1.27
	-0.35	-0.15	-0.43	-0.48	-0.13	-0.07	0.00	-0.08	-0.23	-0.38	0.15	-0.18	-0.24	0.11	-0.18	0.11	0.11	-0.28	-0.24	0.12	-0.12	-0.04	-0.21	0.58	0.68	-0.06	0.53	0.42
	-0.90	-1.17	-1.44	-0.80	-0.76	-0.91	-0.69	-1.04	-1.11	-1.00	-1.09	-1.04	-0.76	-0.94	-0.95	-0.83	-1.23	-1.29	-1.12	-1.00	-1.27	-0.78	-0.98	-0.28	-0.87	-1.34	-0.24	-0.53
<i>C. graminicola</i>	-0.04	-0.02	-0.35	0.93	-0.35	-0.04	0.06	-0.91	-1.47	-0.09	0.27	-0.28	0.63	0.60	0.54	0.66	-0.21	-0.35	-0.37	0.15	-0.16	0.27	0.05	0.32	0.58	0.76	-0.30	-0.33
	0.75	-0.28	0.89	1.20	-0.65	0.27	-0.37	-0.44	-1.91	0.68	0.68	-0.54	0.74	0.66	0.19	0.94	0.86	0.48	0.08	0.64	0.56	0.00	-0.04	-0.02	1.16	0.80	-0.32	0.00
	3.01	0.70	-1.13	2.41	0.18	-1.38	-1.25	-2.37	-0.48	-1.07	-1.38	0.58	0.93	2.16	0.45	-0.26	-1.62	-0.39	-1.02	-0.60	-0.98	0.87	0.23	0.01	-0.66	-1.44	-0.97	-3.34
<i>C. higginsianum</i>	-2.79	-1.10	-1.63	-2.71	0.25	-1.58	-1.32	-2.14	-0.91	-0.63	-1.54	0.14	1.26	-1.97	-0.25	0.60	-1.36	-0.75	-1.07	-1.11	-1.50	-1.07	0.74	0.29	0.70	-2.07	-1.22	-3.78
	-0.89	0.35	0.72	-1.03	1.35	0.98	-0.26	-1.64	-0.27	0.12	0.60	2.11	1.56	-0.42	2.44	1.75	0.37	1.91	1.41	1.00	1.37	0.79	2.23	2.59	1.27	-0.54	-0.22	-1.07
<i>F. radiculosa</i>	0.10	0.53	-1.21	0.11	0.70	-0.11	-0.14	-0.14	-0.27	0.02	0.14	0.07	0.11	-0.41	-0.83	0.65	0.76	0.74	-0.54	0.54	-0.18	0.50	0.08	-0.61	0.03	0.11	-0.58	
	0.26	0.17	0.60	0.49	2.51	-0.45	0.42	0.92	0.96	0.25	-0.66	-0.59	-0.63	3.48	-0.32	-0.43	-0.52	-0.41	0.64	0.82	-0.35	-0.55	-0.77	-0.48	-0.49	-0.15	0.76	-0.19
<i>L. maculans</i>	0.79	0.12	1.42	0.62	3.47	-0.31	-0.71	0.06	1.40	0.88	-0.76	-1.07	-0.96	4.20	-0.89	-0.62	-0.21	-0.20	0.52	0.70	-0.46	-0.73	-0.97	-0.48	-1.16	-0.30	-1.56	0.09
	0.63	0.30	0.97	3.08	-0.17	0.58	1.03	1.11	0.81	1.11	-0.19	-0.47	-0.58	4.36	-0.56	-0.46	-0.40	-0.39	0.91	0.63	-0.20	-0.30	-0.95	-0.53	-1.04	0.23	0.20	0.35
<i>M. oryzae</i>	0.68	0.45	1.22	0.24	-0.27	0.55	-0.03	-0.47	0.29	0.65	-0.20	0.65	0.36	-0.15	0.90	-0.30	0.04	-0.39	0.10	-0.47	-0.30	-0.28	0.21	0.34	-1.20	1.06	-0.07	0.64
	-0.27	-0.40	-0.28	-0.60	-0.06	-0.47	-0.56	-0.43	-0.63	-0.10	-0.52	-1.48	-0.18	0.01	-0.07	0.54	1.04	-0.40	-0.43	-0.54	-0.04	-0.12	-0.09	-0.60	-0.85	-0.58	-0.27	
<i>N. crassa</i>	0.38	0.37	0.38	0.51	0.41	0.31	0.09	0.37	0.84	0.56	0.57	-1.32	0.18	0.43	0.00	-0.18	-0.77	0.08	0.61	0.43	0.73	-0.20	0.24	0.43	-0.41	0.33	0.47	0.31
	-0.18	-0.75	-0.22	0.32	0.14	-0.40	-1.22	-0.35	-0.34	0.35	0.05	0.74	0.09	0.66	-0.30	-0.49	0.19	0.06	-0.78	-0.75	0.13	-0.33	-0.68	0.36	-1.65	-0.70	-0.44	-0.03
	1.02	0.75	1.82	0.65	4.81	0.53	-0.09	1.96	1.00	1.51	-1.19	0.61	0.44	2.14	-0.80	-1.57	-1.44	0.04	0.63	-2.09	0.74	0.59	-0.99	-0.32	-1.04	1.16	-1.19	1.03
	1.72	0.42	0.93	0.86	2.82	-1.08	1.11	-0.48	0.81	-0.49	0.47	-0.51	1.22	-0.63	-0.32	0.13	-0.47	-0.44	-1.45	0.26	0.16	-0.65	-0.73	-0.48	0.03	-0.21	0.01	
	1.05	0.29	1.07	1.15	4.03	-1																						

expression in the condition that was defined as the control in this study. Each line represents a condition of interest, each column a gene of interest (corresponding names of the genes are given in Fig. 6). Genes have been distributed in three groups: the 'R series' that corresponds to 12 candidate reference genes pre-selected from *T. versatilis* data; the 'C series' that corresponds to more classic reference genes previously used in most of gene expression studies, including for filamentous fungi ; and the 'Sc series' that corresponds to genes homologous to *S. cerevisiae* genes, which were previously validated as promising reference genes in this yeast species. Numbers reported in the heat map correspond to $\log_{(\text{base } 2)}$ of FC values. Colour scale and correspondences between $\text{Log}_{(\text{base } 2)}$ and FC values are indicated in the legend (green colour set for a fold-change of 1 ($\log_2 = 0$); red colour arbitrary set for differential expression equal or higher than 12 ($|\log_2| \geq 3,5$). Empty cells: data not available.

Appendix

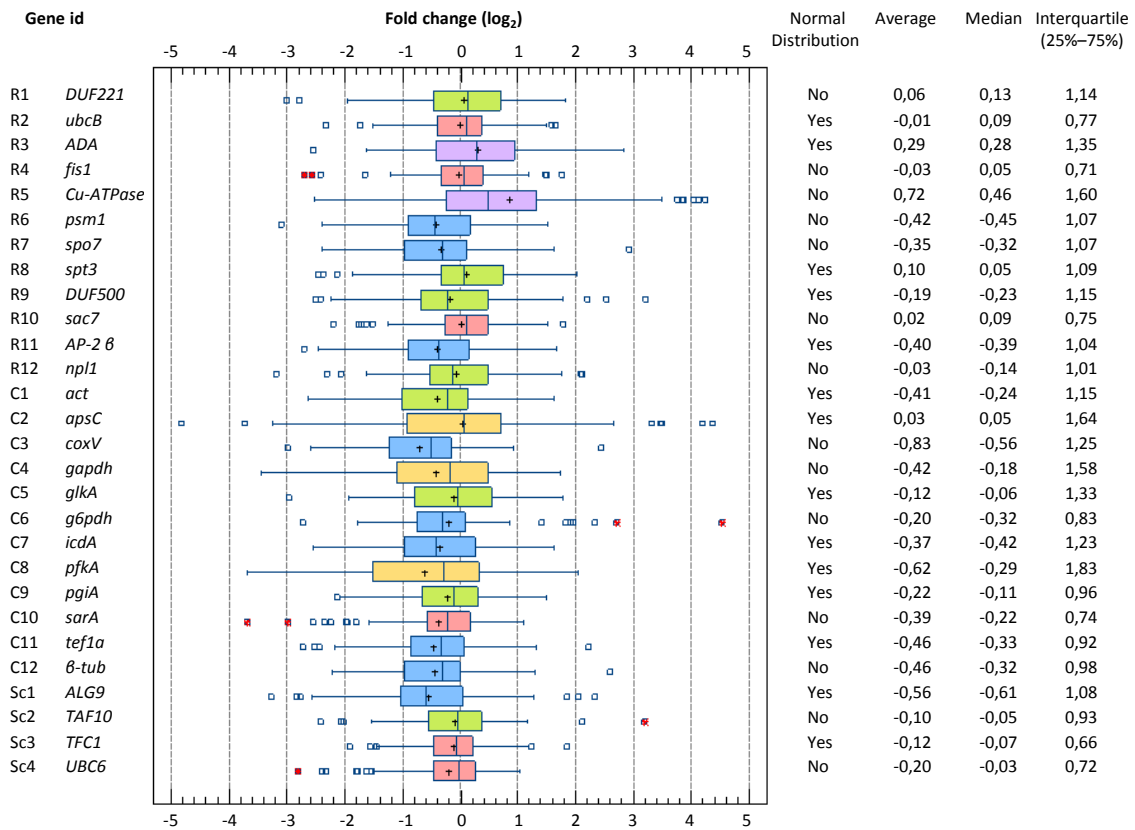


Figure 6: Distribution of fold change values. For each gene, the box-plot gathered about a hundred $\log_2(\text{FC})$ values presented in Fig. 5 (Legend as in Fig. 1). The extreme values amongst outliers are marked with a red asterisk. The colours of the boxes relate the classes that were determined from the HAC (see Fig. 7). From the top to the bottom, we listed the new candidates ('R series'), the classical reference genes ('C series') and the putative references homologous to *S. cerevisiae* genes Sc1–Sc4). The right panel resumes the type of distribution (normal or not), average, median and interquartile for each gene.

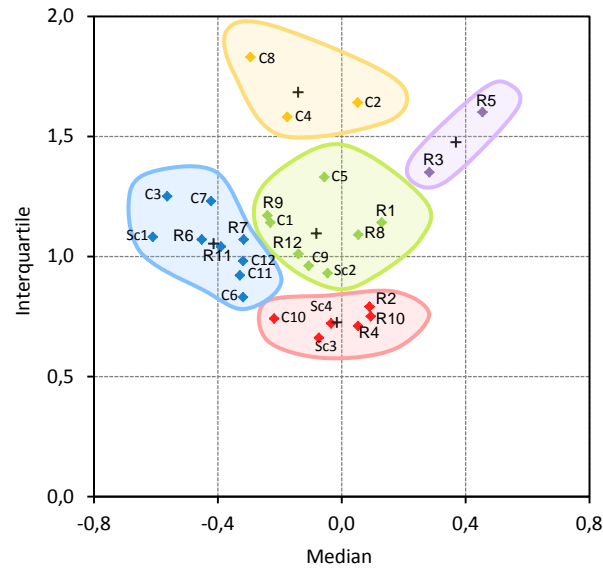


Figure 7: Classification of the reference genes according to their median and interquartile.

Scatter plot of the interquartile *versus* median. The clusters that were obtained by hierarchical classification (HAC, see Additional file 5) are circled with different colours. The black crosses indicate the centroid of each cluster.

Appendix III

Additional File 1: Table of culture conditions.

	Carbon source/ Stress condition	Concentration	Time of exposure
AL1	Glucose	1%	48 hours (inoculum)
AL37	Xylose	0.2%	30 minutes
AL38	Xylose	0.2%	2 hours
AL39	Xylose	1%	1 hour
AL12	Xylose	1%	24hours
AL40	Arabinose	0.2%	30 minutes
AL41	Arabinose	0.2%	2 hours
AL42	Arabinose	1%	1 hour
AL14	Arabinose	1%	24hours
AL2	Cellobiose	0.2%	30 minutes
AL3	Cellobiose	0.2%	2 hours
AL5	Thio-gentiobiose	0.2%	30 minutes
AL6	Thio-gentiobiose	0.2%	2 hours
AL8	Xylobiose	0.2%	30 minutes
AL9	Xylobiose	0.2%	2 hours
AL10	Xylobiose	0.2%	6 hours
AL16	Avicel	1%	24 hours
AL18	Arbocel	1%	24 hours
AL20	Xylan	1%	24 hours
AL22	Wheat straw	1%	24 hours
AL24	Wheat bran	1%	24 hours
AL25	40°C		1 hour
AL26	pH 2		1 hour
AL27	pH 8		1 hour
AL28	KCl	0.5 M	1 hour
AL29	N starvation		1 hour
AL30	C starvation		1 hour
AL31	Conidia on MM + glucose		0
AL32	Conidia on MM + glucose		2 hours
AL33	Conidia on MM + glucose		4 hours
AL34	Conidia on MM + glucose		8 hours
AL35	Conidia on MM + glucose		12 horus
AL36	Conidia on MM + glucose		16 hours

Additional File 2: Gene tags correspondence table.

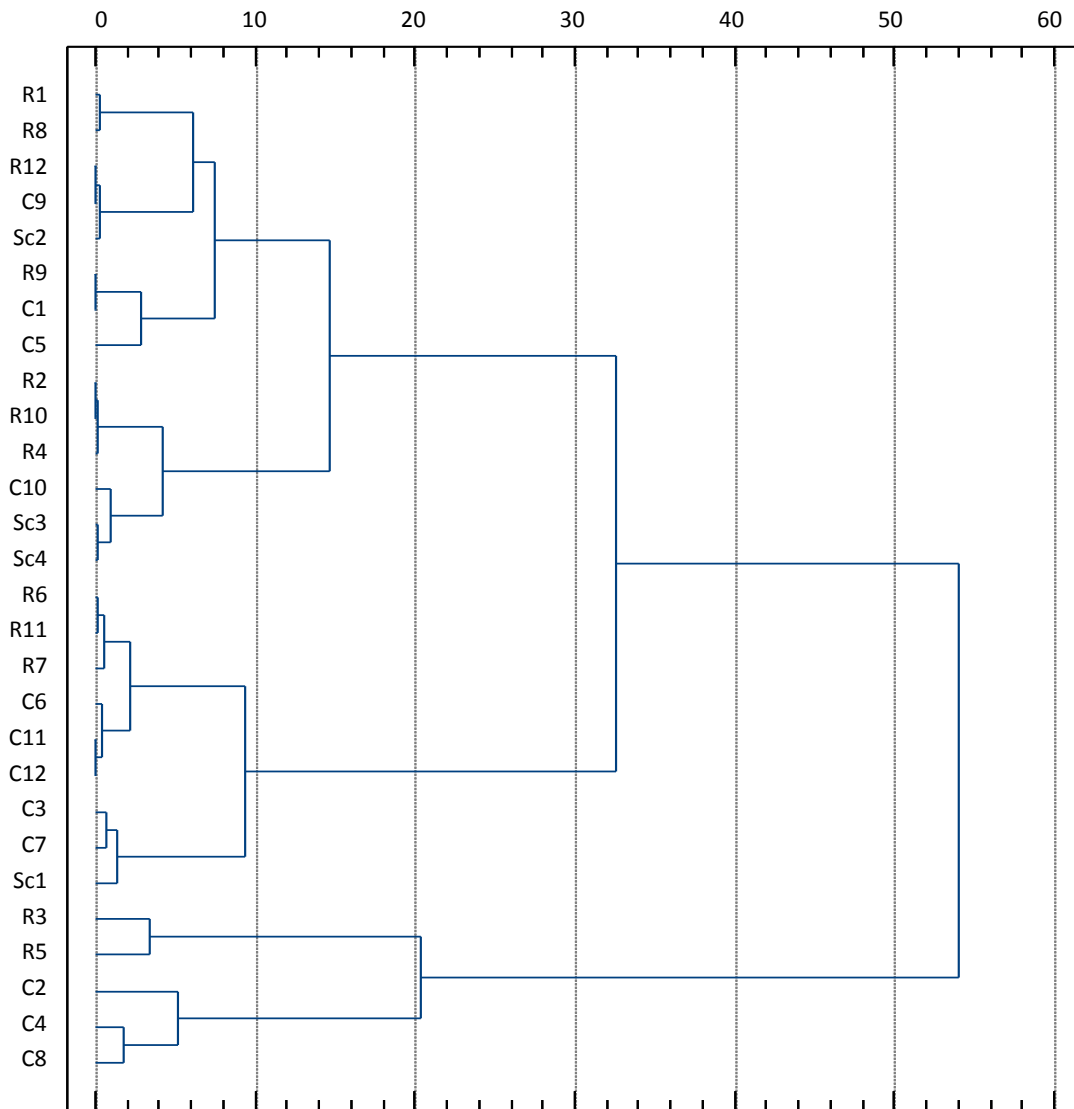
Gene	Abbreviation	Trichoderma reesei	Aspergillus niger	Aspergillus flavus	Aspergillus fumigatus	Aspergillus oryzae	Aspergillus nidulans	Biomberia graminis	Colletotrichum graminicola	Coiletostrichum higginsianum	Fibrospora radiculosa	Magnapotha oryzae	Neurospora crassa Ory74A	Pyrenoma omphalodes	Rhizactonia solani	Sordaria macrospora	Rhizophagus irregularis	Leptostrephus maculans
R1	DUF221	IGeneSH1_pg.C_scaffold_30 p0185	An02g06130	AFLA_081350	AFLA_1G02130A009000500088	AV100556	BgH02039	GLRG_05955	CH063_00871	FIBRA_08801	MGG_08731	NCU00789	PCON_12847	AGI1A08357	SMAC_01470	IGeneSH1_kg.1977_#_2_#_ACT1 GA_L001_R1_(paired)_contig_5_1 208	LEWA_P065510	
R2	ubcB	estEX_GeneWisePlus_C_12 D176	An06g01120	AFLA_008300	AFLA_6G14210A0090010300040	AV8258	BgH06180	GLRG_02867	CH063_00270	FIBRA_06677	MGG_01756	NCU05731	PCON_13196	AGI1A03146	SMAC_04780	IGeneSH1_kg.27642_#_1_#_ste p1_rep_L229	LEWA_P069070	
R3	ADA	e_bw1.1.2.408.1	An08g10280	AFLA_033740	AFLA_6G13180A009000300044	AV5492	BgH00696	GLRG_04176	CH063_01242	FIBRA_06582	MGG_00770	NCU04642	PCON_08457	AGI1A03187	SMAC_07514	CEI161353_2631	LEWA_P092310	
R4	Fis1	estEX_GeneWisePlus_C_scaffold_11 P00055	An02g02190	AFLA_136310	AFLA_2G13320A009002600025	AV6225	BgH01060	GLRG_07224	CH063_15696	FIBRA_02287	MGG_06075	NCU05313	PCON_08457	AGI1A02350	SMAC_01677	CEI52408_803	LEWA_P094420	
R5	Cu-ATPase	estEX_GeneWisePlus_C_20 D194	An01g08110	AFLA_126510	AFLA_1G16130A009003800018	AV3624	BgH05322	GLRG_00559	CH063_09249	FIBRA_04716	MGG_10941	NCU07531	PCON_03959	AGI1A06921	SMAC_09387	IGeneSH1_kg.2532_#_5_#_ACT1 GA_L001_R1_(paired)_contig_9_1 36	LEWA_P077640	
R6	psm1	e_bw1.2.6.118.1	An02g11900	AFLA_086840	AFLA_3G08260A009000500145	AV2963	BgH01540	GLRG_01295	CH063_01410	FIBRA_04207	MGG_03378	NCU01323	PCON_01689	AGI1A04394	SMAC_04013	e_bw1.7.37.4.1	LEWA_P118930	
R7	sp07-like	e_bw1.1.0.307.1	An01g02700	AFLA_028590	AFLA_1G03910A009000300095	AV4045	BgH06743	GLRG_06171	CH063_08642	NF	MGG_06158	NCU06331	PCON_09532	NF	NF	NF	LEWA_P125200	
R8	spt3	estEX_GeneWisePlus_C_20 20297	An08g07230	AFLA_017400	AFLA_1G14030A009001200047	AV0719	BgH02583	GLRG_02875	CH063_05534	FIBRA_02051	MGG_01766	NCU07992	PCON_05320	AGI1A07598	SMAC_01829	IGeneSH1_kg.20346_#_6_#_rem1 bin_c2358	LEWA_P010620	
R9	DUF500	estEX_GeneWisePlus_C_11 345	An17g00750	AFLA_056840	AFLA_5G07310A0090070100022	AV2236	BgH04616	GLRG_02764	CH063_07800	NF	MGG_06145	NCU08844	PCON_09037	AGI1A01139	SMAC_09329	NF	LEWA_P093490	
R10	sac7	estEX_GeneWisePlus_C_20 0093	An10g00490	AFLA_058340	AFLA_2G01060A0090070100037	AV11994	BgH04404	GLRG_00194	CH063_03162	FIBRA_02381	MGG_06390	NCU00196	PCON_02065	AGI1A08941	SMAC_02232	CEI70822_1289	LEWA_P109460	
R11	AP-2 β	estEX_GeneWisePlus_C_20 346	An16g04650	AFLA_044280	AFLA_2G10340A009001100060	AV5950	BgH05612	GLRG_05791	CH063_06823	FIBRA_01138	MGG_04150	NCU05232	PCON_05411	AGI1A01095	SMAC_07169	IGeneSH1_kg.6882_#_2_#_ACT1 GA_L001_R1_(paired)_contig_9_1 043	LEWA_P066650	
R12	np1	estEX_Igenesh1_pm.C_900 64	An01g13070	AFLA_084730	AFLA_1G14940A009000500123	AV0834	BgH01845	GLRG_02179	CH063_11673	FIBRA_08131	MGG_05320	NCU00169	PCON_13312	AGI1A06288	SMAC_02255	IGeneSH1_pg.8724_#_5	LEWA_P089300	
C1	act	estEX_GeneWisePlus_C_11839 1422	An15g00500	AFLA_055230	AFLA_6G04740A0090070100006	AV6542	BgH00992	GLRG_03056	CH063_05085	FIBRA_05357	MGG_03982	NCU04173	PCON_07284	AGI1A08626	SMAC_04416	IGeneSH1_kg.7034_#_1_#_step 3_rep_c1442	LEWA_P099940	
C2	apsC	estEX_pg.C_scaffold_10 P00059	An04g02850	AFLA_097230	AFLA_2G09030A009002000070	AV10198	BgH01996	GLRG_03995	CH063_05712	FIBRA_02224	MGG_09249	NCU03515	PCON_09037	AGI1A05767	SMAC_03210	gim1.14384_β	LEWA_P035690	
C3	coxV	estEX_GeneWisePlus_C_17 D422	An14g04170	AFLA_119650	AFLA_5G10560A009001000048	AV2316	BgH01071	GLRG_00865	CH063_08374	FIBRA_02530	MGG_03188	NCU05457	PCON_06033	AGI1A01681	SMAC_06203	CEI23382_3049	LEWA_P066750	
C4	gapdh	estEX_Igenesh5_pg.C_104 D	An16g01830	AFLA_025100	AFLA_5G01970A009000300132	AV8041	BgH00075	GLRG_05186	CH063_12658	FIBRA_08136	MGG_01084	NCU01528	PCON_12745	AGI1A01681	SMAC_06203	IGeneSH1_kg.371_#_6_#_ACT1 GA_L001_R1_(paired)_contig_91_1 6	LEWA_P066760	
C5	glaA	estEX_GeneWisePlus_C_17 D330	An12g08610	AFLA_130070	AFLA_6G02230A009002000010	AV8689	BgH00115500000101	GLRG_10104	CH063_08332	FIBRA_06174	MGG_03041	NCU00575	PCON_04126	AGI1A04168	SMAC_01265	IGeneSH1_kg.18160_#_1_#_ACT TGA_L001_R1_(paired)_contig_1 3411	LEWA_P038480	
C6	βp9dh	estEX_GeneWisePlus_C_40 239	An02g12140	AFLA_086620	AFLA_3G08470A009000500142	AV2981	BgH00058	GLRG_01682	CH063_10848	FIBRA_04599	MGG_09926	NCU09111	PCON_09179	AGI1A00128	SMAC_04606	gim1.14299_β	LEWA_P021540	
C7	lcaA	estEX_GeneWisePlus_C_27023 7	An02g12430	AFLA_086400	AFLA_3G08660A009000500140	AV2999	BgH00070	GLRG_10180	CH063_09108	FIBRA_06662	MGG_07268	NCU03857	PCON_06644	AGI1A04187	SMAC_06282	IGeneSH1_kg.4624_#_2_#_ACT1 GA_L001_R1_(paired)_contig_9_1 3	LEWA_P001430	
C8	pkfA	estEX_pm.C_scaffold_1 D00425	An18g01670	AFLA_024480	AFLA_4G00960A009000300139	AV3223	BgH00033	GLRG_02587	CH063_04111	FIBRA_02126	MGG_02653	NCU00629	PCON_06425	AGI1A05192	SMAC_01314	IGeneSH1_kg.5818_#_3_#_ACT1 GA_L001_R1_(paired)_contig_8_1 190	LEWA_P074190	
C9	βgA	estEX_pm.C_scaffold_2 P00020	An16g05420	AFLA_044820	AFLA_2G07970A009001000065	AV6037	BgH00028	GLRG_09917	CH063_15722	FIBRA_01802	MGG_12822	NCU07281	PCON_05415	AGI1A02400	SMAC_00816	CE93109_3223	LEWA_P085200	
C10	sara	e_bw1.9.296.1	An01g04040	AFLA_029760	AFLA_1G04940A009000300084	AV0411	BgH01537	GLRG_02181	CH063_02284	FIBRA_06932	MGG_06362	NCU11181	PCON_13389	AGI1A02121	SMAC_06445	CEI49746_2474	LEWA_P093940	
C11	tefla	estEX_GeneWisePlus_C_60370 P2	An18g04840	AFLA_090780	AFLA_1G08390A009002000008	AV4218	BgH01866	GLRG_02486	CH063_14343	FIBRA_04022	MGG_03641	NCU02003	PCON_04930	AGI1A02841	SMAC_05446	IGeneSH1_kg.21690_#_7_#_ACT1 TGA_L001_R1_(paired)_contig_1 213	LEWA_P072560	
C12	β-tub	estEX_Igenesh1_pm.C_400 P2	An01g05180	AFLA_051840	AFLA_1G10910A009003800031	AV1182	BgH00079	GLRG_08786	CH063_01300	FIBRA_02239	MGG_06064	NCU04054	PCON_02618	AGI1A04633	SMAC_00640	Gim1_scaffold_15325-4137- 5956 Beta-tubulin	LEWA_P053950	
Sc1	βlg9	estEX_GeneWisePlus_C_30 472	An08g07020	AFLA_017560	AFLA_1G13870A009001200048	AV10118	BgH00582	GLRG_07260	CH063_01599	FIBRA_00083	MGG_14835	NCU04454	PCON_02596	AGI1A06567	SMAC_01154	IGeneSH1_pm.21462_#_1	LEWA_P122130	
Sc2	TAF10	βw1.32.57.1	An07g07150	AFLA_132710	AFLA_5G11460A009002600068	AV0154	BgH04401	GLRG_06265	CH063_11312	FIBRA_04539	MGG_09574	NCU08729	PCON_06904	AGI1A02354	SMAC_00049	CEI42209_1207	LEWA_P050840	
Sc3	TFCl	e_bw1.7.549.1	An02g10220	AFLA_004750	AFLA_5G02550NF	AV7997	BgH01245	GLRG_04893	CH063_02317	FIBRA_01189	MGG_04019	NCU03787	PCON_08803	AGI1A02354	SMAC_02634	IGeneSH1_pg.14725_#_4	LEWA_P0094510	
Sc4	UBC6	estEX_GeneWisePlus_C_80 789	An14g00230	AFLA_115290	AFLA_4G04610A009002060002	AV10874	BgH06491	GLRG_09476	CH063_03435	FIBRA_04727	MGG_09977	NCU05592	PCON_07434	AGI1A02393	SMAC_03303	IGeneSH1_kg.11151_#_14_#_ste lens_c1105	LEWA_P076760	

Additional File 4: List of classically used reference genes.

		Organism and bibliographic reference									
Gene	Protein	Aspergillus niger - Bohle 2007	Trichoderma reesei - Steiger 2010	Beauveria bassiana - Zhou 2012	Heterobasidi on annosum - Raffaello 2013	Hemileia vastatrix - Vieira 2011	Melampsora larici-populina - Hacquard 2011	Phytophthora parasitica - Zhi 2006	Saccharomyces cerevisiae - Teste 2009		
C1	<i>act</i> Actin	x	x	x	x		x	x			
C2	<i>apsC</i> Aminopeptidase C	x									
C3	<i>coxV</i> Cytochrome C oxidase subunit V	x	x		x	x					
C4	<i>gapdh</i> Glyceraldehyde 3-phosphate dehydrogenase	x		x	x	x	x	x			x
C5	<i>glkA</i> Glucokinase	x	x								
C6	<i>g6pdh</i> Glucose-6-phosphate dehydrogenase	x	x							x	
C7	<i>icdA</i> Isocitrate dehydrogenase precursor	x									
C8	<i>pfkA</i> Phosphofructokinase	x									
C9	<i>pgiA</i> Phosphoglucose isomerase	x									
C10	<i>sarA</i> Secretion associated GTP-binding protein	x	x		x						
C11	<i>tef1a</i> Translation elongation factor 1a		x	x		x			x		x
C12	<i>β-tub</i> Tubulin			x	x	x				x	x

Appendix

Additional File 5: HAC tree.



Appendix IV

Table AIV: Expression values (RPKM) of the differently expressed genes coding for putative transporters, esterases and lipases in Glucose 48h, Straw 24h and Straw + Glucose 5h in *T. versatilis*. Annotations and ORF's I.Ds are also indicated.

ORF	Annotation	RPKM Glucose 48h	RPKM Straw 24h	RPKM Straw + Glucose 5h
Transporters				
Pf170c009g01000	MFS drug efflux pump	1,23	44,73	3,41
Pf160c025g00010	Hexose transporter	3,48	1804,88	13,5
Pf310c008g00100	MFS sugar transporter	9,72	3745,78	130,59
Pf360c035g00060	Maltose permease	2,94	1111,77	24,64
Pf300c043g00040	Maltose permease	3,48	1059,11	8,77
Pf160c049g00400	Small oligopeptide transporter, OPT family	1,73	411,78	17,8
Pf160c064g00610	MFS quinate transporter	0,72	168,62	1,47
Pf360c003g00670	MFS lactose permease	9,58	2226,33	108,28
Pf350c004g02180	MFS transporter	2,4	515,47	6,56
Pf160c064g00160	Major facilitator superfamily transporter	0,84	175,43	2,19
Pf360c036g00140	MFS sugar transporter	1,61	312,66	32,72
Pf300c054g00170	MFS alpha-glucoside transporter	1,66	302,06	3,27
Pf360c038g01260	Oligopeptide transporter	0,7	122,77	12,23
Pf160c049g00760	MFS quinate transporter	1,02	159,06	1,26
Pf360c034g00240	MFS transporter	1,74	240,01	2,82
Pf390c010g00150	Sugar transporter (Hexose transporter)	0,74	84,09	1,76
Pf170c010g02210	MFS glucose transporter	8,54	948,67	65,3
Pf170c005g01030	MFS peptide transporter Ptr2	1,66	162,68	35,3
Pf170c016g02100	Amino acid permease (Dip5)	1,3	124,02	8,95
Pf360c034g00350	Sugar transporter, high affinity	1,15	91,22	1,73
Pf360c034g00470	MFS transporter	1,29	94,27	1,46
Pf360c035g00410	MFS sugar transporter	1,53	110,39	1,61
Pf200c012g01250	ABC multidrug transporter	5,19	351,16	112,21
Pf300c043g00070	Maltose permease	33,49	2216,49	130,01
Pf390c011g00020	Sugar transporter family protein	6,4	419,73	17,1
Pf200c024g00180	MFS transporter	2,5	159,46	36,71
Pf120c036g00090	MFS sugar transporter	1	52,39	4,92
Pf350c004g03830	Sugar transporter	0,58	30,04	0,29
Pf120c031g00510	GABA permease	17,71	890,23	1046,97
Pf120c034g01400	Cell surface receptor/MFS transporter (FLVCR)	2,19	103,4	8,06
Pf120c029g00870	MFS multidrug transporter	2	82,62	21,94
Pf200c038g01050	MFS quinate transporter	1,25	45,25	3,85
Pf300c021g00410	Maltose permease	1,34	46,92	0,85
Pf120c012g00320	MFS sugar transporter protein	4,07	132,34	9,06
Pf120c082g00850	MFS transporter	0,92	29,77	0,54
Pf170c011g01280	2-ketogluconate transporter	2,75	80,44	3,11
Pf160c026g01510	OPT peptide transporter Mtd1	2,73	79,2	14,85
Pf170c012g00360	MFS sugar transporter	1,72	46,04	0,89
Pf360c003g00680	Maltose permease	0,4	10,46	0,68
Pf230c009g00330	Maltose permease	1,8	44,09	2,2
Pf390c013g00590	MFS transporter	2,43	45,82	15,86
Pf120c031g00110	MFS nicotinic acid transporter Tna1	1,48	22,78	2,87

Appendix

Pf390c011g00970	Predicted transporter	0,68	10,36	0,6
Pf230c017g01360	Sugar transporter family protein	2	27,79	1,64
Pf200c020g00360	Amino acid transporter	1,65	19,7	1,21
Pf290c003g00260	MFS alpha-glucoside transporter	1,2	12,76	1,6
Pf120c010g00140	Amino acid permease	1,95	19,38	6,69
Pf200c028g00540	MFS transporter	0,48	2,84	#N/A
Pf200c060g00790	Predicted transporter	0,7	3,73	#N/A
Pf170c016g02250	Monocarboxylate transporter	2,29	10,29	5,96
Pf120c085g00430	Potassium transporter	1,2	3,75	#N/A
Esterases				
Pf300c045g00540	Ferulic acid esterase A Precursor	7,68	3690,18	26,13
Pf160c028g00320	Feruloyl esterase B EC=3.1.1.73 Precursor	10,03	4582,62	26,78
Pf160c026g01540	Acetyl xylan esterase	3,08	1337,97	6,08
Pf200c039g00480	Acetyl xylan esterase (Axe1)	7,13	2669,19	10,98
Pf300c043g00050	Feruloyl esterase	1,94	483,86	3,05
Pf230c025g00020	Rhamnogalacturonan acetylerase RgaE	1,66	383,83	3,02
Pf390c013g00750	Acetylxylan esterase 2	11,43	1988,52	168,25
Pf360c038g00960	Pectinesterase EC=3.1.1.11	0,85	169,99	5,21
Pf160c015g00140	Carboxylesterase	1,35	218,3	3,93
Pf160c010g00040	Carboxylesterase	0,6	74,71	0,7
Pf120c019g01440	Phosphoesterase superfamily protein	0,63	59,46	4,8
Pf200c060g01100	Pectin methylesterase	0,43	35,53	0,7

Appendix V

Table AV: Expression values (RPKM) of all *T. versatilis* GH-encoding genes in Glucose 48h, Straw 24h and Straw + Glucose 5h. Differentially expressed (DE) genes are indicated in the last column. Annotations, corresponding protein CAZy family number and ORF's I.Ds are also indicated.

Enzyme	CAZy family	ORF	RPKM Glucose 48h	RPKM Straw 24h	RPKM Straw + Glucose 5h	DE
β-glucosidase	GH1	Pf120c012g00340	1,01	123,17	1,37	Yes
		Pf120c019g00490	13,34	1052,94	113,38	Yes
		Pf160c025g02130	2,27	2,88	2,86	No
		Pf270c010g00330	1,33	53,78	1,7	Yes
β-galactosidase	GH2	Pf170c010g00800	2,7	24,43	16,61	No
		Pf200c019g00200	1,45	4,39	1,04	No
β-mannosidase	GH2	Pf120c029g00070	10,66	30,29	7,04	No
		Pf120c037g00830	1,47	18,04	4,78	No
		Pf170c016g02380	1,61	19,67	11,21	Yes
		Pf340c003g00500	0,26	0,47	0,44	No
		Pf350c004g03000	2	207,6	18,44	Yes
		Pf390c011g00980	0,93	19,86	4,69	Yes
β-glucuronidase	GH2	Pf240c018g00060	0,48	0,78	0,61	No
Exo-β-d-glucosaminidase	GH2	Pf170c006g00250	3,09	504,91	7,92	Yes
β-glucosidase	GH3	Pf120c028g00030	13,39	7,47	4,6	No
		Pf120c035g00170	1,57	0,93	0,77	No
		Pf120c082g02430	2,68	191,59	14,43	Yes
		Pf160c026g00890	0,65	0,81	0,39	No
		Pf160c027g00430	2,25	559,61	6,17	Yes
		Pf170c008g00300	16,93	9,64	15,41	No
		Pf170c012g00610	7,82	183,26	79,83	Yes
		Pf200c024g00260	2,36	8,19	1,72	No
		Pf200c029g00930	0,27	0,08	0,48	No
		Pf200c038g00070	0,28	4,08	0,95	Yes
		Pf300c021g00940	0,55	18,93	0,43	Yes
		Pf300c046g01230	2,23	2,12	1,74	No
		Pf330c021g00740	0,85	14,75	0,87	Yes
		Pf350c004g03120	4,08	409,73	65,57	Yes
		Pf350c004g03380	2,2	0,85	1,06	No
		Pf350c007g00270	1,76	3,01	1,5	No
		Pf360c035g00420	0,48	30,5	5,67	Yes
		Pf370c002g00210	2,05	662,02	9,59	Yes
		Pf390c012g00150	56,94	0,78	34,25	No
		Pf350c004g03370	1,22	0,44	0,28	No
β-xylosidase	GH3	Pf300c048g00360	5,15	1860,03	8,32	Yes
		Pf350c004g00340	0,74	259,59	2,9	No
		Pf160c028g00250	22,68	23,74	17,12	No
β-N-acetylhexosaminidase	GH3	Pf170c008g00220	8,23	6,32	7,26	No
		Pf340c003g00780	6,24	219,07	36,15	Yes
		Pf120c012g00820	0,57	59,54	0,77	Yes
β-1,4-endoglucanase	GH5	Pf120c019g00180	3,75	1522,4	5,55	Yes

Appendix

		Pf200c012g00180	10,22	4786,47	109,67	Yes
		Pf390c012g00080	6,13	2367,22	26,73	Yes
β-1,4- endomannanase		Pf120c037g00790	0,16	0,96	1,06	No
		Pf330c014g00330	3,09	654,48	7,65	Yes
Exo-1,3-β- glucanase		Pf160c017g00350	36,27	29,3	41,55	No
		Pf200c029g00140	0,33	0,38	0,61	No
Endo-1,6-β- glucosidase		Pf120c035g00410	0,86	5,36	1,13	Yes
Cellobiohydrolas e (non reducing end)	GH6	Pf200c030g01540	9,78	4175,04	71,09	Yes
Cellobiohydrolas e (reducing end)	GH7	Pf200c030g01760	0,56	19,2	2,19	Yes
		Pf360c030g01900	44,59	18674,4	161,05	Yes
β-1,4- endoglucanase		Pf390c010g00580	6,04	2241,78	105,45	Yes
β-1,4- endoxyfanase	GH10	Pf340c013g00740	7,33	3365,24	37,24	Yes
β-1,4- endoxyfanase	GH11	Pf120c031g00040	28,15	13591,11	72,95	Yes
		Pf160c017g00290	3,11	433,47	24,47	Yes
		Pf160c028g00460	8,93	2733,75	15,83	Yes
		Pf170c009g01150	3,35	1454,47	14,8	Yes
		Pf300c053g00050	0,74	72,51	2,53	No
		Pf330c014g01020	0,4	15,59	0,66	No
Xyloglucan-active β-1,4- endoglucanase	GH12	Pf390c014g00360	69,52	28354,08	306,13	Yes
		Pf160c027g00440	2,19	1258,07	4,17	Yes
α-amylase	GH13	Pf360c034g00360	45,74	134,52	10,21	No
		Pf120c013g00040	0,5	41,83	35,43	No
		Pf120c027g00060	1,32	5,05	3,28	No
		Pf230c019g00420	46,75	11,46	18,99	No
		Pf350c004g01270	0,89	3,36	0,93	No
		Pf350c004g03400	0,37	5,71	4,08	No
		Pf350c014g00980	0,56	52,07	1,89	Yes
		Pf360c030g01060	0,57	10,8	5,88	No
α-1,6-glucosidase	GH13	Pf360c030g01080	0,62	11,64	6,48	No
		Pf160c049g00840	16,83	21,68	17,09	No
α-1,3-glucan synthase	GH13	Pf350c004g01050	0,52	0,68	0,48	No
		Pf160c027g00050	0,86	0,94	0,88	No
Glycogen debranching enzyme	GH13	Pf300c055g00790	0,72	6,58	19,98	No
		Pf120c029g01300	64,1	38,53	97,91	No
Glucoamylase	GH15	Pf300c021g00880	0,4	0,19	3,68	No
		Pf360c030g00830	0,51	1,78	0,84	No
		Pf360c030g01070	1,37	127,92	13,45	Yes
Chitinase	GH18	Pf230c019g00100	8,02	3204,3	13,13	Yes
Endomannanase	GH26	Pf120c029g00740	0,86	165,1	1,65	Yes
α-galactosidase	GH27	Pf120c082g00840	2,84	1356,54	22,85	Yes
		Pf160c026g01530	2,04	945,34	9,74	Yes
		Pf160c028g00270	1,37	21,99	1,68	Yes
		Pf350c004g02720	1,72	0,83	1,83	No
		Pf360c030g00650	0,68	78,29	1,28	Yes
Endopolygalactur	GH28	Pf120c033g00190	2,18	1,27	1,77	No

onase		Pf260c002g00060	1,16	357,6	7,91	Yes
		Pf350c004g02860	0,61	1,77	1,14	No
		Pf360c030g01930	0,38	93,43	8,71	No
		Pf300c046g00050	70,56	45,72	79,99	No
		Pf360c003g00200	0,43	0,16	0,52	No
Endorhamnogalacturonase		Pf120c010g00150	4,08	1246,21	8,07	Yes
		Pf170c010g00950	0,53	0,2	0,32	No
		Pf200c024g00240	0,79	162,53	4,34	Yes
		Pf230c008g00080	0,37	7,62	0,69	No
		Pf360c030g00660	0,32	0,57	0,35	No
Exopolygalacturonase		Pf160c022g00090	0,64	0,5	0,7	No
		Pf170c012g00120	2,43	84,87	3,21	Yes
		Pf200c039g00740	0,38	1,03	0,72	No
		Pf390c011g02100	0,81	152,36	9,43	No
Xylogalacturonan hydrolase		Pf350c004g00270	0,98	3,76	2,01	No
α -l-fucosidase	GH29	Pf120c012g01410	1,09	191,07	1,49	No
		Pf230c017g00370	0,86	6,86	1,81	No
α -glucosidase	GH31	Pf120c027g00010	1,49	66,37	1,71	Yes
		Pf120c082g02120	183,83	105,62	223,38	No
		Pf170c005g00530	210,24	30,17	11,24	No
		Pf170c015g00030	3,45	12,32	3,9	No
		Pf300c021g00340	1,05	285,52	2,76	Yes
		Pf350c004g01200	5,27	3,36	3,01	No
		Pf350c013g00830	0,38	0,48	12,55	No
		Pf350c014g00820	3,38	313,95	33,01	Yes
		Pf350c014g01000	0,79	27,09	1,3	Yes
		Pf360c038g00680	22,41	71,03	12,8	No
α -xylosidase		Pf230c009g00220	1,2	45,96	17,27	Yes
Endoinulinase	GH32	Pf300c021g00870	0,36	0,39	0,65	No
Exoinulinase		Pf170c009g00080	0,56	0,17	0,29	No
		Pf300c021g00860	0,31	0,34	0,48	No
Invertase		Pf200c030g00520	0,65	2,33	0,15	No
		Pf350c004g01060	0,61	0,62	0,46	No
		Pf120c082g01420	0,2	0,4	0,36	No
		Pf300c021g00770	0,56	0,68	0,53	No
β -galactosidase	GH35	Pf120c019g01110	5,94	2278,82	41,6	Yes
		Pf160c025g01170	0,54	8,08	1,18	No
		Pf160c049g00650	0,43	27,78	0,63	No
		Pf160c049g00920	1,12	252,23	1,6	Yes
		Pf390c013g00430	1,01	102,41	9,05	Yes
α -galactosidase	GH36	Pf120c028g00090	1,25	255,88	1,19	Yes
		Pf300c046g01460	0,55	0,76	0,75	No
		Pf360c003g00500	6,74	2,14	2,22	No
		Pf360c003g00520	0,99	86,51	2,06	Yes
α -l-arabinofuranosidase	GH43	Pf160c026g01520	8,18	795,37	8,61	Yes
Endoarabinase		Pf120c084g00520	0,15	0,32	0,32	No
		Pf160c028g00400	0,43	115,86	1,33	Yes
		Pf330c014g00370	0,67	3,37	3,31	No
		Pf360c003g00410	0,34	0,35	0,6	No

Appendix

β-xylosidase		Pf200c061g00170	2,43	591,35	5,23	Yes
		Pf300c045g00020	4,58	1115,68	19,38	Yes
		Pf350c004g02750	0,25	0,46	0,69	No
		Pf360c036g00150	0,67	14,74	1,06	Yes
		Pf360c034g00340	1,04	240,73	1,93	Yes
β-1,4-xylosidase/α-arabinofuranosidase		Pf100c010g01370	2,98	29,51	4,73	No
		Pf120c035g00110	0,52	2,51	0,79	No
β-1,4-endoglucanase	GH45	Pf200c038g00680	9,43	3737,37	97,56	Yes
		Pf340c013g00130	8,25	103,89	132,29	No
α-l-arabinofuranosidase	GH51	Pf120c036g00100	4,06	24,49	8,79	No
		Pf120c084g00540	0,21	0,09	0,34	No
Endo-1,4-β-galactanase	GH53	Pf330c021g00580	1,52	502,45	4,21	Yes
α-l-arabinofuranosidase	GH54	Pf170c016g00470	1,39	377,44	7,55	Yes
		Pf200c012g01410	8,61	3344,02	16,11	Yes
		Pf360c038g00430	0,83	181,75	2,73	Yes
		Pf160c025g02760	0,46	12,48	1,22	No
		Pf100c010g00180	0,75	6,42	1,23	Yes
β-1,4-endoglucanase	GH61	Pf230c019g00440	0,7	0,49	0,76	No
		Pf350c013g00110	7,6	756,84	9,35	Yes
Arabinoxylan arabinofuranohydrolase	GH62	Pf160c017g00280	0,75	64,76	6,47	Yes
		Pf170c008g00480	3,26	1659,62	21,32	Yes
		Pf390c013g00450	3,03	1074,15	12,68	Yes
α-glucuronidase	GH67	Pf120c034g01610	2,64	5,63	1,97	No
		Pf300c043g00030	2,93	147,52	2,22	Yes
		Pf330c022g00050	25,81	125,27	14,03	No
α-1,3-glucanase	GH71	Pf300c058g00790	3,01	1381,44	8,69	Yes
Endoglucanase	GH74	Pf120c031g00120	5,91	2496,6	42,09	Yes
α-l-rhamnosidase	GH78	Pf100c002g00030	1,6	4,2	1,91	No
		Pf100c002g00040	1,01	0,73	0,98	No
		Pf160c028g00330	2,14	156,46	3,38	Yes
		Pf170c012g00110	0,38	85,37	1,76	Yes
		Pf300c046g01210	0,36	0,31	0,42	No
		Pf300c055g00510	9,21	3,57	8,76	No
		Pf350c004g03800	1,82	10,94	2,38	No
d-4,5 unsaturated glucuronyl hydrolase	GH88	Pf170c006g00030	0,6	2,96	0,36	Yes
		Pf390c010g00160	3,13	73,05	4,68	Yes

Appendix VI

Table AVI: Expression values (RPKM) of all *T. versatilis* transcription factor-encoding genes in Glucose 48h, Straw 24h and Straw + Glucose 5h. ORF's I.Ds and annotations are also indicated.

ORF	Annotation	RPKM Glucose 48h	RPKM Straw 24h	RPKM Straw + Glucose 5h	Known TF
Pf230c011g00330	C2H2 type conidiation transcription factor BrIA	0,26	27,01	42,07	brIA
Pf120c085g00070	C6 finger domain protein	2,48	208,57	59,85	
Pf240c022g00550	C2H2 finger domain protein	1,96	133,59	59,77	
Pf200c012g01220	C2H2 finger domain protein	0,81	40,44	23,92	
Pf200c080g00170	C6 finger domain protein	3,96	167,85	101,32	
Pf360c003g00510	C6 transcription factor,	1,68	60,6	2,59	
Pf170c010g00810	C6 transcription factor	1,52	49,12	6,86	crlA
Pf200c038g01030	Nitrogen assimilation transcription factor nit-4	0,98	28,92	0,96	
Pf260c002g00100	Putative Zn(II)2Cys6 transcription factor (Eurofung)	2,25	49,08	19,82	
Pf390c010g00180	Fungal specific transcription factor	4,61	88,28	3,17	
Pf160c024g00040	C6 zinc finger domain protein	0,52	9,84	0,7	alcR
Pf160c009g00080	C6 transcription factor	7,54	119,69	12,31	
Pf160c026g00520	C6 transcription factor	10,77	168,13	19,5	xlnR
Pf170c006g00050	C6 transcription factor	1,05	12,8	0,59	
Pf390c012g00040	C2H2 finger domain protein	1,83	15,28	4,78	

Appendix VII

Table AVII: Antisense expression values (RPKM) of all *T. versatilis* genes containing NATs (natural antisense transcripts) in Glucose 48h, Straw 24h and Straw + Glucose 5h. ORF I.Ds and annotations are also indicated.

ORF	Annotation	RPKM Glucose 48h	RPKM Straw 24h	RPKM Straw + Glucose 5h
Pf390c011g01990	hypothetical protein	2258,71	1265,65	1874,49
Pf160c024g00010	hypothetical protein	1121,76	962,74	1032,92
Pf170c007g00100	hypothetical protein	71,11	81,60	78,04
Pf170c009g00750	hypothetical protein	56,44	39,99	68,19
Pf160c024g00220	hypothetical protein	31,26	73,79	63,41
Pf200c030g01190	hypothetical protein	50,79	73,06	60,35
Pf120c012g01010	hypothetical protein	13,37	15,80	40,74
Pf350c004g00230	hypothetical protein	35,68	14,09	39,04
Pf300c048g00190	hypothetical protein	55,23	17,33	37,33
Pf330c021g00690	hypothetical protein	14,94	33,69	33,94
Pf330c014g00430	hypothetical protein	18,41	40,29	27,95
Pf120c078g00860	hypothetical protein	25,06	8,36	27,78
Pf300c055g01230	hypothetical protein	37,25	39,12	25,85
Pf160c064g00720	hypothetical protein	14,78	36,65	24,22
Pf300c055g00660	hypothetical protein	35,86	20,71	24,18
Pf150c002g00050	hypothetical protein	13,38	17,25	23,90
Pf160c027g00930	hypothetical protein	15,60	30,99	23,72
Pf350c003g00010	hypothetical protein	30,79	23,25	23,09
Pf120c085g00080	hypothetical protein	24,63	7,23	19,29
Pf120c038g00110	hypothetical protein	38,28	18,07	19,14
Pf160c027g00740	hypothetical protein	39,69	8,53	17,81
Pf300c041g00070	hypothetical protein	17,57	26,66	15,66
Pf360c038g00050	hypothetical protein	18,16	1,93	14,81
Pf390c011g01230	hypothetical protein	17,85	10,72	13,32
Pf160c027g01190	hypothetical protein	17,81	13,07	12,42
Pf120c078g00320	hypothetical protein	5,20	17,33	12,03
Pf200c021g00080	hypothetical protein	13,18	3,53	11,99
Pf120c082g01630	hypothetical protein	13,26	5,03	10,94
Pf230c024g00050	hypothetical protein	13,76	7,52	10,83
Pf350c019g00100	hypothetical protein	23,07	3,07	10,71
Pf200c030g01390	hypothetical protein	9,61	8,29	10,50
Pf300c007g00010	hypothetical protein	19,58	5,55	10,48
Pf160c018g00310	hypothetical protein	8,18	0,57	10,43
Pf160c025g00370	hypothetical protein	17,78	2,74	10,25
Pf200c060g00950	hypothetical protein	10,67	2,06	10,08
Pf160c020g00090	hypothetical protein	5,70	2,31	10,05
Pf250c001g00040	hypothetical protein	30,36	38,42	9,68
Pf170c011g00800	hypothetical protein	11,45	3,15	9,61
Pf170c016g00630	hypothetical protein	25,99	5,32	9,43
Pf160c015g00370	hypothetical protein	15,47	11,40	9,37
Pf120c012g00220	hypothetical protein	5,02	10,02	7,26
Pf390c002g00010	hypothetical protein	7,94	0,51	7,16

Appendix

Pf160c027g00110	hypothetical protein	7,24	10,25	7,15
Pf210c008g00060	hypothetical protein	7,23	4,52	6,77
Pf160c025g02330	hypothetical protein	8,71	8,23	6,54
Pf330c013g00020	hypothetical protein	7,03	3,91	6,53
Pf390c017g00130	hypothetical protein	13,98	14,77	6,50
Pf170c008g00360	hypothetical protein	21,65	1,66	5,67
Pf350c013g00890	hypothetical protein	6,96	1,35	5,40
Pf200c059g00770	hypothetical protein	5,02	2,31	5,15
Pf160c017g00320	hypothetical protein	8,14	4,09	5,01
Pf300c056g00310	hypothetical protein	5,89	2,06	4,66
Pf330c014g00360	hypothetical protein	7,92	5,13	4,64
Pf180c013g00010	hypothetical protein	12,14	3,59	4,53
Pf300c041g00060	hypothetical protein	5,64	2,89	4,07
Pf300c010g00010	hypothetical protein	9,61	2,35	3,43
Pf390c013g00010	hypothetical protein	6,51	8,23	3,40
Pf120c034g01060	hypothetical protein	5,96	2,50	3,39
Pf120c036g01010	hypothetical protein	5,11	4,85	3,37
Pf290c006g00020	hypothetical protein	8,96	0,16	3,23
Pf150c018g00030	hypothetical protein	14,31	1,01	3,18
Pf300c048g00250	hypothetical protein	6,95	5,20	2,89
Pf120c037g00020	hypothetical protein	7,11	2,47	2,75
Pf370c002g00670	hypothetical protein	10,00	1,22	1,93
Pf150c019g00020	hypothetical protein	5,08	0,69	1,91
Pf250c007g00010	hypothetical protein	8,33	0,00	0,21
Pf390c010g00130	hypothetical protein	5,08	0,00	0,16
Pf300c055g00230	hypothetical protein	10,24	0,00	0,11
Pf180c007g00020	hypothetical protein	5,77	0,20	0,11
Pf370c001g00060	hypothetical protein	12,06	0,21	0,00
Pf200c012g00480	hypothetical protein	9,81	0,07	0,00
Pf200c038g01070	hypothetical protein	0,31	5,62	12,24
Pf170c011g00720	hypothetical protein	4,18	7,14	7,95
Pf200c030g01400	hypothetical protein	2,41	5,30	7,61
Pf160c002g00040	hypothetical protein	2,30	5,09	7,20
Pf230c025g00330	hypothetical protein	1,08	9,28	6,24
Pf160c026g00260	hypothetical protein	2,71	15,69	5,97
Pf350c014g00050	hypothetical protein	2,75	7,35	5,86
Pf120c034g00010	hypothetical protein	3,39	7,59	5,64
Pf170c011g01480	hypothetical protein	0,15	7,73	5,55
Pf160c001g00060	hypothetical protein	4,54	5,08	5,46
Pf100c010g01230	hypothetical protein	4,54	5,23	4,49
Pf300c021g00040	hypothetical protein	4,80	15,88	4,33
Pf120c029g01240	hypothetical protein	4,34	7,63	3,39
Pf350c010g00020	hypothetical protein	2,29	8,23	2,66
Pf170c010g01990	hypothetical protein	1,34	5,41	2,65
Pf160c049g00950	hypothetical protein	2,54	5,03	2,42
Pf160c008g00400	hypothetical protein	3,95	5,03	2,24
Pf170c016g00460	hypothetical protein	0,00	5,71	1,78
Pf340c026g00050	hypothetical protein	2,46	22,90	1,56
Pf160c028g00020	hypothetical protein	3,00	6,14	1,03
Pf230c024g00040	hypothetical protein	1,47	25,65	0,69
Pf300c021g00650	hypothetical protein	0,17	5,53	0,15
Pf310c008g00090	hypothetical protein	0,26	5,48	0,00

Pf160c027g00250	hypothetical protein	2,08	3,37	7,75
Pf120c034g00430	hypothetical protein	1,10	2,95	5,80
Pf360c003g00620	Aldo/keto reductase	1,95	12,44	4,32
Pf160c003g00010	C2H2 finger domain protein	5,25	2,96	2,20
Pf300c055g00500	C6 transcription factor	88,79	58,92	52,41
Pf200c038g01250	Catalytic activity: D-arabinitol + NAD	1,44	10,30	1,35
Pf200c039g00490	Endo-1,4-beta-xylanase 1	0,37	6,73	0,25
Pf330c015g00580	Oligopeptide transporter, putative	2,36	43,19	11,53
Pf300c011g00010	Pc16g15520 protein Fragment - <i>Penicillium chrysogenum</i>	13,08	1,58	0,79
Pf390c020g00050	Pc17g01040 protein Fragment - <i>Penicillium chrysogenum</i>	15,46	21,98	19,92
Pf300c020g00230	Pc20g10680 protein - <i>Penicillium chrysogenum</i>	99,09	50,11	67,22
Pf240c019g00190	Pc20g10680 protein - <i>Penicillium chrysogenum</i>	87,98	59,93	60,36
Pf360c030g00900	Pc20g10680 protein - <i>Penicillium chrysogenum</i>	1250,27	1184,81	1249,53
Pf240c019g00360	Pc20g10680 protein - <i>Penicillium chrysogenum</i>	113,46	91,41	84,19
Pf120c036g00590	Pc20g10680 protein - <i>Penicillium chrysogenum</i>	104,66	60,41	77,47
Pf160c018g00010	Pc20g10680 protein - <i>Penicillium chrysogenum</i>	17,23	8,23	7,57
Pf300c018g00010	Predicted protein	250,19	156,97	180,54
Pf330c012g00040	Purine nucleoside phosphorylase I, inosine and guanosine-specific	27,31	7,49	21,18
Pf360c003g00130	Putative uncharacterized protein	243,07	4,57	12,77
Pf230c028g00040	Putative uncharacterized protein	6,04	4,19	4,80
Pf300c015g00040	Putative uncharacterized protein	41,01	16,83	27,59
Pf120c078g01040	Putative uncharacterized protein	1,84	5,48	1,33
Pf370c002g00890	Putative uncharacterized protein	7,60	4,96	4,58
Pf160c025g02400	Putative uncharacterized protein	9,98	4,53	5,01
Pf120c082g01370	Putative uncharacterized protein	56,02	4,79	13,93
Pf120c013g00420	Putative uncharacterized protein	19,79	6,70	5,90
Pf170c010g02540	Putative uncharacterized protein	5,61	5,66	3,22
Pf120c037g01180	Putative uncharacterized protein	21,35	7,07	5,66
Pf240c017g00140	Putative uncharacterized protein	8,56	3,94	8,12
Pf200c060g00920	Putative uncharacterized protein	4,74	7,78	5,96
Pf200c060g00970	Putative uncharacterized protein	6,59	1,88	4,18
Pf230c004g00040	Putative uncharacterized protein	5,57	2,89	4,60
Pf360c034g00110	Putative uncharacterized protein	5,64	5,22	3,73
Pf160c025g01060	Putative uncharacterized protein	11,26	6,38	7,24
Pf340c025g00040	Putative uncharacterized protein	7,04	8,36	5,92
Pf150c014g00030	Putative uncharacterized protein	16,90	29,07	19,73
Pf300c023g00010	Putative uncharacterized protein	9,00	5,28	9,87
Pf390c010g00120	Putative uncharacterized protein	9,72	0,22	3,54
Pf210c002g00010	Putative uncharacterized protein	21,99	2,89	1,92
Pf300c020g00390	Putative uncharacterized protein	7,52	0,75	2,43
Pf200c080g00010	Putative uncharacterized protein	6,97	0,93	0,62
Pf240c005g00130	Putative uncharacterized protein Pc13g01090 - <i>Penicillium chrysogenum</i>	7,43	8,62	6,19
Pf390c011g00520	Putative uncharacterized protein Pc21g20800 - <i>Penicillium chrysogenum</i>	6,70	8,90	5,85
Pf170c011g01660	Quercetin 2,3-dioxygenase	0,07	22,23	0,18
Pf350c012g00620	Quinone oxidoreductase	8,31	0,95	1,21
Pf170c011g00130	Tricarboxylate transport protein (Ctp)	4,97	5,89	1,67
Pf240c019g00260	3-hydroxyanthranilate 3,4-dioxygenase Bna1	5,75	2,77	2,76
Pf120c036g00160	3-oxoacyl-(Acyl-carrier-protein) reductase -	4,06	6,68	2,72
Pf230c017g00630	3'-phosphoadenosine-5'-phosphosulfate reductase Met16	5,48	5,65	3,53
Pf160c028g00190	4-hydroxyphenylpyruvate dioxygenase	11,08	6,37	5,72
Pf120c034g02150	50S ribosomal protein L2	5,92	12,10	11,72
Pf170c005g00680	Acetyl-CoA-acetyltransferase	5,68	2,97	3,61

Appendix

Pf160c025g01730	Actin monomer binding protein	5,27	2,67	4,89
Pf300c058g00350	Acyl-CoA dehydrogenase	7,47	1,47	6,53
Pf160c014g00790	Acylphosphatase EC=3.6.1.7	60,06	84,73	70,43
Pf100c010g00800	ADP-ribosylation factor family protein	29,97	40,18	25,41
Pf160c026g00550	Aldehyde dehydrogenase family protein	5,10	0,76	3,25
Pf330c021g01170	Aldo/keto reductase	2,25	11,70	4,45
Pf120c033g00270	Aldo-keto reductase (AKR)	0,00	5,74	0,00
Pf120c085g00490	Aldo-keto reductase	15,26	2,71	4,08
Pf160c021g00310	Alpha-1,3-mannosyltransferase (Alg3)	7,89	2,33	5,86
Pf330c016g00650	Amino acid permease family protein	5,57	1,37	0,51
Pf350c004g02070	APSES transcription factor Xbp1	6,59	3,72	3,96
Pf100c010g00790	Aspergillopepsin-2	2,32	5,93	0,53
Pf160c009g00440	Beta-1,6-glucan boisynthesis protein (Knh1)	9,74	0,98	9,60
Pf160c003g00020	C2H2 finger domain protein	29,49	18,26	14,37
Pf160c021g00400	C2H2 transcription factor	7,52	1,51	1,40
Pf350c006g00980	C6 transcription factor	3,34	5,31	2,42
Pf200c024g00190	Calmodulin-binding protein Sha1	0,04	6,43	0,75
Pf120c079g00470	Carbonyl reductase	0,93	12,10	1,89
Pf350c006g00110	Cation diffusion facilitator 1	12,60	5,39	11,38
Pf160c001g00040	Cation diffusion facilitator family metal ion transporter	5,08	1,71	3,54
Pf170c009g00720	Chromosome segregation protein Cse1	5,02	0,39	0,95
Pf360c030g01140	Class I alpha-mannosidase	7,10	0,83	8,21
Pf330c016g01080	Ctr copper transporter family protein	8,11	4,76	6,91
Pf350c004g03440	Cytochrome P450	0,33	8,66	0,29
Pf330c014g01190	Dhp1-interacting protein Din1	9,49	4,44	5,41
Pf120c013g00570	Dienelactone hydrolase family protein	17,08	20,73	25,74
Pf390c012g00580	Dienelactone hydrolase family protein	2,77	1,43	5,07
Pf360c030g00440	DNA repair protein Rad26	507,29	137,66	278,48
Pf170c016g00380	DNA-3-methyladenine glycosylase	26,47	2,97	13,23
Pf270c010g00910	DUF1212 domain membrane protein	0,27	5,49	0,58
Pf120c019g01650	DUF1275 domain protein	7,90	1,31	4,63
Pf120c019g00610	DUF718 domain protein	4,79	6,09	1,56
Pf160c021g00190	DUF985 domain protein	6,91	7,83	3,04
Pf160c027g01400	Endonuclease/exonuclease/phosphatase family protein	7,10	6,54	8,73
Pf350c012g00260	Esterase/lipase	0,40	3,75	5,05
Pf120c067g00270	Eukaryotic translation initiation factor 6	0,06	6,40	0,80
Pf160c049g00550	FAD dependent oxidoreductase superfamily -	5,93	1,66	1,36
Pf170c009g02860	Fructosyl amino acid oxidasesarcosine oxidase	15,72	0,11	0,08
Pf230c009g00490	FYVE domain protein	9,54	1,00	1,96
Pf120c081g00070	Gamma-glutamyltranspeptidase	6,31	2,17	11,96
Pf160c049g00560	Glutathione-dependent formaldehyde-activating enzyme	26,84	3,07	16,00
Pf120c036g00700	Glycerol-3-phosphate phosphatase (GppA)	7,32	0,66	1,44
Pf170c014g00040	Haloacid dehalogenase, type II	1510,69	518,05	883,66
Pf100c002g00990	HEAT repeat protein	7,07	2,03	5,93
Pf120c035g00530	Heat shock protein/chaperonin HSP78	106,28	109,33	133,22
Pf200c059g00010	Hemoglobin and proliferation regulated protein	5,99	2,85	5,30
Pf330c014g00760	Histone acetyltransferase complex component Epl1	15,37	7,63	11,48
Pf120c078g00910	HIT domain protein	3,35	22,38	2,90
Pf170c009g02060	LEA domain protein	11,01	6,00	9,93
Pf150c019g00590	Lipase/esterase	6,35	4,28	4,05
Pf300c048g00280	Malic enzyme	1,32	5,69	0,99
Pf230c017g01160	MFS sugar transporter	5,18	3,22	3,55

Pf120c036g00090	MFS sugar transporter	10,92	4,70	5,60
Pf240c005g00110	Mitochondrial chaperone bcs1	7,77	3,12	1,95
Pf120c037g01160	Mitochondrial DnaJ chaperone (Tim14)	14,47	8,69	5,88
Pf120c013g00720	Myb-like DNA-binding domain protein	59,94	2,86	3,30
Pf350c004g01390	N,O-diacetyl muramidase	5,12	2,42	2,97
Pf390c014g00020	Nuclear transport factor 2 domain protein	7,10	11,44	2,82
Pf370c002g01180	Nucleolar RNase III	7,36	5,19	3,53
Pf350c013g00040	Oleate delta-12 desaturase	0,00	5,17	0,06
Pf160c049g00840	Oligo-1,6-glucosidase	6,56	0,49	1,69
Pf390c014g00090	Oxidoreductase, short chain dehydrogenase/reductase family superfamily	161,90	94,62	98,06
Pf330c012g00160	Oxidoreductase, short-chain dehydrogenase/reductase family	5,92	3,75	1,97
Pf170c009g02500	Pathogenesis associated protein Cap20	6,23	40,58	9,31
Pf160c001g00020	strong similarity to P. Pc06g02340 protein - <i>Penicillium chrysogenum</i>	254,01	140,96	77,05
Pf300c018g00100	strong similarity to P. Pc13g08550 protein - <i>Penicillium chrysogenum</i>	7,81	1,78	3,76
Pf340c013g00720	strong similarity to P. Pc16g09500 protein Precursor - <i>Penicillium chrysogenum</i>	0,08	12,22	0,26
Pf170c010g02490	strong similarity to P. Pc22g03400 protein - <i>Penicillium chrysogenum</i>	21,65	4,60	14,90
Pf300c046g00650	strong similarity to P. Pc22g16180 protein - <i>Penicillium chrysogenum</i>	0,08	5,79	1,15
Pf120c019g00360	Peptidase D, putative	6,22	1,03	8,25
Pf170c009g00630	Peroxisomal D3,D2-enoyl-CoA isomerase	12,78	2,54	5,78
Pf350c013g00380	PfkB family carbohydrate kinase (Mak32)	13,27	3,46	8,73
Pf160c025g01750	PHD and RING finger domain protein	10,92	1,28	3,26
Pf300c046g01180	Phosphatase	5,93	3,74	4,78
Pf120c033g00370	Phosphate transporter (Pho88)	6,18	1,00	8,19
Pf230c017g01350	Phosphoglycerate mutase family protein	1,51	5,47	2,50
Pf340c013g00100	Predicted protein	19,50	6,39	1,68
Pf170c011g01730	Predicted protein	2,46	1,12	5,28
Pf160c027g00120	Protein RTM1, putative	9,52	6,34	4,11
Pf170c009g00470	Protein transport protein sec22	7,36	2,60	4,67
Pf120c037g01340	Putative uncharacterized protein	11,41	2,43	9,53
Pf350c012g00160	Putative uncharacterized protein	2,35	5,66	3,44
Pf170c010g01210	Putative uncharacterized protein	4,51	4,20	6,02
Pf230c017g01400	Putative uncharacterized protein	4,16	5,48	2,93
Pf120c082g00390	Putative uncharacterized protein	3,51	3,72	7,94
Pf200c025g00270	Putative uncharacterized protein	5,80	4,23	3,28
Pf200c020g00490	Putative uncharacterized protein	7,57	6,71	5,33
Pf360c038g01250	Putative uncharacterized protein	0,20	9,37	2,01
Pf120c033g00800	Putative uncharacterized protein	5,62	0,81	2,94
Pf160c027g00210	Putative uncharacterized protein	49,04	4,15	4,79
Pf200c029g00990	Putative uncharacterized protein	19,23	14,88	7,52
Pf390c013g00220	Putative uncharacterized protein	1,89	10,39	2,69
Pf170c009g02120	Putative uncharacterized protein	1,01	17,05	0,97
Pf200c029g00650	Putative uncharacterized protein	9,02	2,14	3,02
Pf350c012g00230	Putative uncharacterized protein	16,09	1,38	3,16
Pf120c083g00050	Putative uncharacterized protein	0,12	8,18	0,16
Pf120c033g01250	Putative uncharacterized protein	0,82	14,05	0,53
Pf120c037g00380	Putative uncharacterized protein	8,15	0,68	9,12
Pf120c032g00060	Putative uncharacterized protein	4,44	8,10	1,84
Pf390c010g00440	Putative uncharacterized protein	4,86	5,68	2,35
Pf300c058g01120	Putative uncharacterized protein	9,94	3,59	8,77
Pf230c017g00470	Putative uncharacterized protein	16,53	0,38	2,27
Pf230c019g00110	Putative uncharacterized protein	0,00	12,24	0,28

Appendix

Pf360c033g00500	Putative uncharacterized protein	8,51	8,41	5,74
Pf120c037g00580	Putative uncharacterized protein	1,05	10,71	0,71
Pf350c012g00360	Putative uncharacterized protein	7,79	11,67	9,05
Pf200c030g01470	Putative uncharacterized protein	7,74	1,39	4,69
Pf360c003g00460	Putative uncharacterized protein	0,57	6,39	0,09
Pf120c082g00920	Putative uncharacterized protein	0,05	11,83	0,30
Pf120c029g00010	Putative uncharacterized protein	13,43	0,68	0,26
Pf170c010g01430	Putative uncharacterized protein	6,77	18,75	14,46
Pf170c009g01060	Putative uncharacterized protein	20,02	19,79	17,51
Pf300c054g00840	Putative uncharacterized protein	3,11	16,26	2,35
Pf200c012g00560	Putative uncharacterized protein	1,47	5,08	3,05
Pf100c010g00360	Putative uncharacterized protein	13,85	4,97	6,62
Pf360c030g00210	Putative uncharacterized protein	6,77	1,62	3,61
Pf160c025g00860	Putative uncharacterized protein	9,08	4,69	2,97
Pf170c010g00140	Putative uncharacterized protein	10,39	5,48	5,09
Pf160c015g00110	Putative uncharacterized protein	5,96	0,95	1,50
Pf160c027g00230	Putative uncharacterized protein	285,08	19,63	63,65
Pf230c003g00170	Putative uncharacterized protein	2,52	6,64	1,65
Pf160c025g01540	Putative uncharacterized protein	144,63	5,64	5,34
Pf270c010g00720	Putative uncharacterized protein	523,08	189,57	303,22
Pf120c082g00930	Putative uncharacterized protein	5,36	1,93	4,02
Pf360c030g00580	Putative uncharacterized protein	5,24	4,58	5,01
Pf160c026g00870	Putative uncharacterized protein	28,52	0,87	1,06
Pf370c002g00330	Putative uncharacterized protein	5,63	2,40	2,96
Pf240c004g00100	Putative uncharacterized protein	6,72	1,12	4,70
Pf300c048g00390	Putative uncharacterized protein	12,22	5,31	12,12
Pf200c030g00960	Putative uncharacterized protein	6,45	4,67	4,89
Pf230c003g00050	Putative uncharacterized protein	11,49	7,28	5,11
Pf120c085g00110	Putative uncharacterized protein	5,91	11,00	5,37
Pf170c013g01490	Putative uncharacterized protein	36,47	5,92	8,98
Pf170c006g00200	Putative uncharacterized protein	3,70	28,55	2,72
Pf360c038g00900	Putative uncharacterized protein	4,77	3,90	5,41
Pf230c003g00030	Putative uncharacterized protein	5,15	11,68	3,47
Pf200c029g01130	Putative uncharacterized protein	6,90	2,91	2,68
Pf100c002g00910	Putative uncharacterized protein	44,56	3,75	0,17
Pf300c047g00020	Putative uncharacterized protein	20,89	1,30	6,30
Pf120c079g00560	Putative uncharacterized protein	4,86	7,74	13,92
Pf100c002g00730	Putative uncharacterized protein	8,32	5,98	6,00
Pf350c013g00670	Putative uncharacterized protein	15,46	8,41	11,55
Pf160c025g01640	Putative uncharacterized protein	6,15	6,31	7,06
Pf330c012g00230	Putative uncharacterized protein	11,45	3,92	7,12
Pf390c010g00590	Putative uncharacterized protein	0,00	5,08	0,45
Pf390c014g00030	Putative uncharacterized protein	21,45	10,74	14,88
Pf160c015g00360	Putative uncharacterized protein	2,81	5,21	3,11
Pf230c017g00950	Pyridoxamine phosphate oxidase	43,98	0,35	0,04
Pf120c012g00890	Pyruvate kinase EC=2.7.1.40	4,17	6,13	3,51
Pf160c014g00080	Remark: N-terminally truncated due to contig border. Fragment	11,69	2,82	7,36
Pf120c034g02090	RWD domain protein (Gir2)	5,30	1,04	2,83
Pf200c038g00760	SAP domain protein	0,29	21,33	5,60
Pf120c034g00160	Secreted glucosidase	0,14	13,75	0,22
Pf160c026g00830	Short chain dehydrogenase/oxidoreductase	5,99	3,30	3,15
Pf300c045g00010	Short-chain dehydrogenase	0,06	10,68	0,15

Pf390c010g00620	Short-chain dehydrogenase	29,59	9,20	17,48
Pf200c060g00600	Short-chain dehydrogenase/reductase 2	0,55	4,30	12,50
Pf200c019g00350	Short-chain dehydrogenase/reductase family protein	6,89	5,12	4,27
Pf170c009g02580	Short-chain dehydrogenase/reductase family protein	11,45	1,79	9,31
Pf160c026g01570	Short-chain dehydrogenases/reductase	0,04	12,80	0,17
Pf360c030g00570	SYF2 splicing factor family protein	28,87	0,37	6,48
Pf120c082g02080	Thymidylate kinase	8,40	3,85	6,34
Pf170c015g00070	Translation initiation factor IF-2	10,17	1,78	4,18
Pf160c026g01420	tRNA splicing endonuclease subunit (Sen54)	122,30	61,13	77,26
Pf170c016g00080	U3 small nucleolar ribonucleoprotein protein Mpp10	4,31	8,51	10,60
Pf300c058g00860	YeeE/YedE family integral membrane protein	6,16	3,85	4,04
Pf170c010g01500	Zinc-binding oxidoreductase, putative	5,34	3,73	4,97
Pf300c058g01290	Methylthioribose-1-phosphate isomerase EC=5.3.1.23	9,82	6,22	6,35
Pf160c028g00040	FAD linked oxidase domain protein	13,97	6,11	12,21
Pf150c017g00040	Putative uncharacterized protein	6,45	1,10	0,76
Pf150c019g00540	Putative uncharacterized protein	10,17	8,10	9,67

Appendix VIII

Table AVIII: Normalised expression values of the GOI in *T.versatilis* $\Delta xlnR$ exposed to different carbon sources. The standard error of the mean (SEM) is also shown.

Gene	Normalised expression	Samples									
		AL1	AL18	AL20	AL22	AL29	AL30	AL38	AL39	AL41	AL42
GOI2	Expression	1,00	15544,8	6039,10	77347,0	0,46	0,79	0,82	0,71	2,05	0,74
	SEM	0,03	106,47	32,46	1803,91	0,02	0,07	0,06	0,06	0,10	0,05
GOI4	Expression	1,00	2610,61	699,53	11104,9	1,12	0,96	0,90	0,74	0,68	0,69
	SEM	0,04	39,55	4,98	244,25	0,10	0,03	0,06	0,02	0,07	0,06
GOI6	Expression	1,00	41,43	248,22	883,81	1,08	0,19	4,55	0,54	0,37	0,23
	SEM	0,04	1,57	1,51	18,94	0,03	0,02	0,16	0,06	0,04	0,03
GOI7	Expression	1,00	0,79	20,38	27,24	0,80	0,71	64,80	4,60	0,65	1,04
	SEM	0,02	0,01	0,21	1,07	0,03	0,03	1,82	0,11	0,01	0,02
GOI10	Expression	1,00	239,52	66,66	411,80	0,62	2,19	2,79	1,70	2,71	0,97
	SEM	0,02	2,35	0,79	9,61	0,03	0,19	0,07	0,07	0,04	0,03
GOI12	Expression	1,00	1,23	2,26	3,54	1,02	1,01	1,64	1,46	1,20	1,41
	SEM	0,04	0,03	0,07	0,14	0,05	0,05	0,08	0,10	0,06	0,07
GOI14	Expression	1,00	534,87	683,37	6932,74	1,28	2,76	126,34	24,77	3,81	1,46
	SEM	0,02	7,95	6,92	139,30	0,04	0,22	4,06	1,58	0,39	0,07
GOI15	Expression	1,00	78,50	9,09	134,57	1,29	1,14	3,47	1,30	3,98	0,78
	SEM	0,03	0,81	0,07	2,75	0,02	0,03	0,14	0,04	0,05	0,02
GOI19	Expression	1,00	32,18	58,05	45,17	0,87	3,13	112,22	18,95	10,50	3,87
	SEM	0,02	0,35	2,25	0,91	0,08	0,25	3,23	0,60	0,26	0,16
GOI21	Expression	1,00	11,59	17,48	122,41	0,82	1,23	2,30	1,24	1,48	1,01
	SEM	0,05	0,40	0,31	2,89	0,05	0,10	0,19	0,03	0,06	0,09
GOI23	Expression	1,00	3,10	10,53	16,95	0,79	11,23	78,02	33,63	9,82	2,69
	SEM	0,02	0,03	0,40	0,37	0,05	0,25	3,70	0,95	0,27	0,04
GOI26	Expression	1,00	17,25	13,23	318,51	0,54	0,53	2,09	0,75	1,28	0,52
	SEM	0,04	0,79	0,38	10,86	0,05	0,04	0,13	0,03	0,07	0,04
GOI27	Expression	1,00	1,10	6,58	9,91	1,19	4,15	1,47	5,62	1,10	2,32
	SEM	0,02	0,01	0,23	0,21	0,02	0,18	0,04	0,19	0,04	0,04
GOI30	Expression	1,00	96,30	21,93	266,96	1,36	3,63	2,19	2,16	1,73	1,24
	SEM	0,02	2,37	0,42	10,33	0,06	0,20	0,08	0,09	0,03	0,03
GOI31	Expression	1,00	3,81	7,53	50,52	0,38	1,72	0,82	0,66	0,40	0,60
	SEM	0,11	0,15	0,37	2,12	0,05	0,04	0,08	0,06	0,02	0,02
GOI33	Expression	1,00	2409,26	474,58	7614,73	1,22	1,21	1,44	0,98	0,73	0,68
	SEM	0,02	18,53	2,53	172,72	0,04	0,08	0,12	0,07	0,01	0,03

References

- Adams, T. H., Wieser, J. K. & Yu, J. H. (1998).** Asexual sporulation in *Aspergillus nidulans*. *Microbiol Mol Biol Rev MMBR* **62**, 35–54.
- Anders, S. & Huber, W. (2010).** Differential expression analysis for sequence count data. *Genome Biol* **11**, R106.
- Andersen, C. L., Jensen, J. L. & Ørntoft, T. F. (2004).** Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res* **64**, 5245–5250.
- Andersen, M. R., Salazar, M. P., Schaap, P. J., van de Vondervoort, P. J. I., Culley, D., Thykaer, J., Frisvad, J. C., Nielsen, K. F., Albang, R. & other authors. (2011).** Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88. *Genome Res* **21**, 885–897.
- Aragona, M., Minio, A., Ferrarini, A., Valente, M. T., Bagnaresi, P., Orrù, L., Tononi, P., Zamperin, G., Infantino, A. & other authors. (2014).** De novo genome assembly of the soil-borne fungus and tomato pathogen *Pyrenochaeta lycopersici*. *BMC Genomics* **15**, 313.
- Archer, D. B. & Turner, G. (2006).** Genomics of Protein Secretion and Hyphal Growth in *Aspergillus*. In *Fungal Genomics, The Mycota*, pp. 75–96. Edited by P. D. A. J. P. Brown. Springer Berlin Heidelberg.
- Archer, D. B., Connerton, I. F. & MacKenzie, D. A. (2008).** Filamentous fungi for production of food additives and processing aids. *Adv Biochem Eng Biotechnol* **111**, 99–147.
- Aro, N., Saloheimo, A., Ilmén, M. & Penttilä, M. (2001).** ACEII, a novel transcriptional activator involved in regulation of cellulase and xylanase genes of *Trichoderma reesei*. *J Biol Chem* **276**, 24309–24314.
- Aro, N., Ilmén, M., Saloheimo, A. & Penttilä, M. (2003).** ACEI of *Trichoderma reesei* is a repressor of cellulase and xylanase expression. *Appl Environ Microbiol* **69**, 56–65.

References

- Aro, N., Pakula, T. & Penttilä, M. (2005).** Transcriptional regulation of plant cell wall degradation by filamentous fungi. *FEMS Microbiol Rev* **29**, 719–739.
- Arthanari, Y., Heintzen, C., Griffiths-Jones, S. & Crosthwaite, S. K. (2014).** Natural Antisense Transcripts and Long Non-Coding RNA in *Neurospora crassa*. *PLoS ONE* **9**, e91353.
- Arvas, M., Pakula, T., Lanthaler, K., Saloheimo, M., Valkonen, M., Suortti, T., Robson, G. & Penttilä, M. (2006).** Common features and interesting differences in transcriptional responses to secretion stress in the fungi *Trichoderma reesei* and *Saccharomyces cerevisiae*. *BMC Genomics* **7**, 32.
- Baker, S. C., Bauer, S. R., Beyer, R. P., Brenton, J. D., Bromley, B., Burrill, J., Causton, H., Conley, M. P., Elespuru, R. & other authors. (2005).** The External RNA Controls Consortium: a progress report. *Nat Methods* **2**, 731–734.
- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M. & other authors. (2013).** NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**, D991–995.
- Battaglia, E., Visser, L., Nijssen, A., van Veluw, G. J., Wösten, H. A. B. & de Vries, R. P. (2011a).** Analysis of regulation of pentose utilisation in *Aspergillus niger* reveals evolutionary adaptations in Eurotiales. *Stud Mycol* **69**, 31–38.
- Battaglia, E., Hansen, S. F., Leendertse, A., Madrid, S., Mulder, H., Nikolaev, I. & de Vries, R. P. (2011b).** Regulation of pentose utilisation by AraR, but not XlnR, differs in *Aspergillus nidulans* and *Aspergillus niger*. *Appl Microbiol Biotechnol* **91**, 387–397.
- Battaglia, E., Klaubauf, S., Vallet, J., Ribot, C., Lebrun, M.-H. & de Vries, R. P. (2013).** Xlr1 is involved in the transcriptional control of the pentose catabolic pathway, but not hemi-cellulolytic enzymes in *Magnaporthe oryzae*. *Fungal Genet Biol FG B* **57**, 76–84.
- Van den Berg, M. A., Albang, R., Albermann, K., Badger, J. H., Daran, J.-M., Driessen, A. J. M., Garcia-Estrada, C., Fedorova, N. D., Harris, D. M. & other authors. (2008).** Genome

- sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. *Nat Biotechnol* **26**, 1161–1168.
- Bey, M., Zhou, S., Poidevin, L., Henrissat, B., Coutinho, P. M., Berrin, J.-G. & Sigoillot, J.-C. (2013).** Cello-oligosaccharide oxidation reveals differences between two lytic polysaccharide monoxygenases (family GH61) from *Podospora anserina*. *Appl Environ Microbiol* **79**, 488–496.
- Blackwell, M. (2011).** The Fungi: 1, 2, 3 ... 5.1 million species? *Am J Bot* **98**, 426–438.
- Bloom, J. S., Khan, Z., Kruglyak, L., Singh, M. & Caudy, A. A. (2009).** Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics* **10**, 221.
- Bohle, K., Jungebloud, A., Göcke, Y., Dalpiaz, A., Cordes, C., Horn, H. & Hempel, D. C. (2007).** Selection of reference genes for normalisation of specific gene quantification data of *Aspergillus niger*. *J Biotechnol* **132**, 353–358.
- Boraston, A. B., Bolam, D. N., Gilbert, H. J. & Davies, G. J. (2004).** Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J* **382**, 769–781.
- Brakhage, A. A. & Schroeckh, V. (2011).** Fungal secondary metabolites – Strategies to activate silent gene clusters. *Fungal Genet Biol*, Special Issue: Fungal Secondary Metabolism **48**, 15–22.
- Bridge & Buddie. (2013).** *Talaromyces versatilis* - Index Fungorum - Names Record.
- Brunner, K., Lichtenauer, A. M., Kratochwill, K., Delic, M. & Mach, R. L. (2007).** Xyr1 regulates xylanase but not cellulase formation in the head blight fungus *Fusarium graminearum*. *Curr Genet* **52**, 213–220.
- Bustin, S. A., Benes, V., Nolan, T. & Pfaffl, M. W. (2005).** Quantitative real-time RT-PCR--a perspective. *J Mol Endocrinol* **34**, 597–601.
- Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W. & other authors. (2009).** The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* **55**, 611–622.

References

- Caffall, K. H. & Mohnen, D. (2009).** The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. *Carbohydr Res* **344**, 1879–1900.
- Calero-Nieto, F., Di Pietro, A., Roncero, M. I. G. & Hera, C. (2007).** Role of the transcriptional activator XlnR of *Fusarium oxysporum* in regulation of xylanase genes and virulence. *Mol Plant-Microbe Interact MPMI* **20**, 977–985.
- Cerqueira, G. C., Arnaud, M. B., Inglis, D. O., Skrzypek, M. S., Binkley, G., Simison, M., Miyasato, S. R., Binkley, J., Orvis, J. & other authors. (2014).** The *Aspergillus* Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. *Nucleic Acids Res* **42**, D705–710.
- Chen, S., Xu, J., Liu, C., Zhu, Y., Nelson, D. R., Zhou, S., Li, C., Wang, L., Guo, X. & other authors. (2012).** Genome sequence of the model medicinal mushroom *Ganoderma lucidum*. *Nat Commun* **3**, 913.
- Chen, X., Ishida, N., Todaka, N., Nakamura, R., Maruyama, J., Takahashi, H. & Kitamoto, K. (2010).** Promotion of Efficient Saccharification of Crystalline Cellulose by *Aspergillus fumigatus* Swo1. *Appl Environ Microbiol* **76**, 2556–2561.
- Chikamatsu, G., Shirai, K., Kato, M., Kobayashi, T. & Tsukagoshi, N. (1999).** Structure and expression properties of the endo-beta-1,4-glucanase A gene from the filamentous fungus *Aspergillus nidulans*. *FEMS Microbiol Lett* **175**, 239–245.
- Coradetti, S. T., Craig, J. P., Xiong, Y., Shock, T., Tian, C. & Glass, N. L. (2012).** Conserved and essential transcription factors for cellulase gene expression in ascomycete fungi. *Proc Natl Acad Sci U S A* **109**, 7397–7402.
- Coradetti, S. T., Xiong, Y. & Glass, N. L. (2013).** Analysis of a conserved cellulase transcriptional regulator reveals inducer-independent production of cellulolytic enzymes in *Neurospora crassa*. *MicrobiologyOpen* **2**, 595–609.

- Cuomo, C. A., Güldener, U., Xu, J.-R., Trail, F., Turgeon, B. G., Di Pietro, A., Walton, J. D., Ma, L.-J., Baker, S. E. & other authors. (2007). The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science* **317**, 1400–1402.
- Dagenais, T. R. T. & Keller, N. P. (2009). Pathogenesis of *Aspergillus fumigatus* in Invasive Aspergillosis. *Clin Microbiol Rev* **22**, 447–465.
- Dang, Y., Yang, Q., Xue, Z. & Liu, Y. (2011). RNA interference in fungi: pathways, functions, and applications. *Eukaryot Cell* **10**, 1148–1155.
- Dashtban, M., Schraft, H. & Qin, W. (2009). Fungal bioconversion of lignocellulosic residues; opportunities & perspectives. *Int J Biol Sci* **5**, 578–595.
- Deacon, J. W. (2006). *Fungal Biology*. Wiley.
- Dean, R. A., Talbot, N. J., Ebbole, D. J., Farman, M. L., Mitchell, T. K., Orbach, M. J., Thon, M., Kulkarni, R., Xu, J.-R. & other authors. (2005). The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature* **434**, 980–986.
- Delmas, S., Pullan, S. T., Gaddipati, S., Kokolski, M., Malla, S., Blythe, M. J., Ibbett, R., Campbell, M., Liddell, S. & other authors. (2012). Uncovering the Genome-Wide Transcriptional Responses of the Filamentous Fungus *Aspergillus niger* to Lignocellulose Using RNA Sequencing. *PLoS Genet* **8**.
- Delmas, S., Llanos, A., Parrou, J.-L., Kokolski, M., Pullan, S. T., Shunburne, L. & Archer, D. B. (2014). Development of an Unmarked Gene Deletion System for the Filamentous Fungi *Aspergillus niger* and *Talaromyces versatilis*. *Appl Environ Microbiol* AEM.00625–14.
- Denton, J. A. & Kelly, J. M. (2011). Disruption of *Trichoderma reesei cre2*, encoding an ubiquitin C-terminal hydrolase, results in increased cellulase activity. *BMC Biotechnol* **11**, 103.
- Derveaux, S., Vandesompele, J. & Hellemans, J. (2010). How to do successful gene expression analysis using real-time PCR. *Methods, The ongoing Evolution of qPCR* **50**, 227–230.
- Devonshire, A. S., Elaswarapu, R. & Foy, C. A. (2010). Evaluation of external RNA controls for the standardisation of gene expression biomarker measurements. *BMC Genomics* **11**, 662.

References

- Dheda, K., Huggett, J. F., Chang, J. S., Kim, L. U., Bustin, S. A., Johnson, M. A., Rook, G. A. W. & Zumla, A. (2005).** The implications of using an inappropriate reference gene for real-time reverse transcription PCR data normalization. *Anal Biochem* **344**, 141–143.
- Die, J. V. & Román, B. (2012).** RNA quality assessment: a view from plant qPCR studies. *J Exp Bot* **63**, 6069–6077.
- Dodd, D. & Cann, I. K. O. (2009).** Enzymatic deconstruction of xylan for biofuel production. *Glob Change Biol Bioenergy* **1**, 2–17.
- Donaldson, M. E. & Saville, B. J. (2013).** *Ustilago maydis* natural antisense transcript expression alters mRNA stability and pathogenesis. *Mol Microbiol* **89**, 29–51.
- Donaldson, M. E. & Saville, B. J. (2012).** Natural antisense transcripts in fungi. *Mol Microbiol* **85**, 405–417.
- Dorak, M. T. (2007).** *Real-time PCR*. Garland Science.
- Edgar, R., Domrachev, M. & Lash, A. E. (2002).** Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207–210.
- Erdeniz, N., Mortensen, U. H. & Rothstein, R. (1997).** Cloning-free PCR-based allele replacement methods. *Genome Res* **7**, 1174–1183.
- Espino-Rammer, L., Ribitsch, D., Przulucka, A., Marold, A., Greimel, K. J., Herrero Acero, E., Guebitz, G. M., Kubicek, C. P. & Druzhinina, I. S. (2013).** Two novel class II hydrophobins from *Trichoderma* spp. stimulate enzymatic hydrolysis of poly(ethylene terephthalate) when expressed as fusion proteins. *Appl Environ Microbiol* **79**, 4230–4238.
- Faccioli, P., Ciceri, G. P., Provero, P., Stanca, A. M., Morcia, C. & Terzi, V. (2007).** A combined strategy of ‘in silico’ transcriptome analysis and web search engine optimization allows an agile identification of reference genes suitable for normalization in gene expression studies. *Plant Mol Biol* **63**, 679–688.
- Faghihi, M. A. & Wahlestedt, C. (2009).** Regulatory roles of natural antisense transcripts. *Nat Rev Mol Cell Biol* **10**, 637–643.

- Fang, W. & Bidochka, M. J. (2006).** Expression of genes involved in germination, conidiogenesis and pathogenesis in *Metarhizium anisopliae* using quantitative real-time RT-PCR. *Mycol Res* **110**, 1165–1171.
- Fernandez, J. & Wilson, R. A. (2014).** Cells in cells: morphogenetic and metabolic strategies conditioning rice infection by the blast fungus *Magnaporthe oryzae*. *Protoplasma* **251**, 37–47.
- Fernandez, J., Wright, J. D., Hartline, D., Quispe, C. F., Madayiputhiya, N. & Wilson, R. A. (2012).** Principles of Carbon Catabolite Repression in the Rice Blast Fungus: Tps1, Nmr1-3, and a MATE-Family Pump Regulate Glucose Metabolism during Infection. *PLoS Genet* **8**, e1002673.
- Fierro, F., Laich, F., García-Rico, R. O. & Martín, J. F. (2004).** High efficiency transformation of *Penicillium nalgiovense* with integrative and autonomously replicating plasmids. *Int J Food Microbiol* **90**, 237–248.
- Filer, K. (2003).** Industrial Production of Enzymes for the Feed Industry. In *New Horiz Biotechnol*, pp. 1–16. Edited by S. Roussos, C. R. Soccol, A. Pandey & C. Augur. Springer Netherlands.
- Floudas, D., Binder, M., Riley, R., Barry, K., Blanchette, R. A., Henrissat, B., Martínez, A. T., Otilar, R., Spatafora, J. W. & other authors. (2012).** The Paleozoic Origin of Enzymatic Lignin Decomposition Reconstructed from 31 Fungal Genomes. *Science* **336**, 1715–1719.
- Frandsen, R. J. N. (2011).** A guide to binary vectors and strategies for targeted genome modification in fungi using *Agrobacterium tumefaciens*-mediated transformation. *J Microbiol Methods* **87**, 247–262.
- Frisvad, J. C., Yilmaz, N., Thrane, U., Rasmussen, K. B., Houbraken, J. & Samson, R. A. (2013).** *Talaromyces atroroseus*, a new species efficiently producing industrially relevant red pigments. *PLoS One* **8**, e84102.
- Fullwood, M. J., Wei, C.-L., Liu, E. T. & Ruan, Y. (2009).** Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res* **19**, 521–532.

References

- Galagan, J. E., Calvo, S. E., Borkovich, K. A., Selker, E. U., Read, N. D., Jaffe, D., FitzHugh, W., Ma, L.-J., Smirnov, S. & other authors. (2003).** The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422**, 859–868.
- Galagan, J. E., Calvo, S. E., Cuomo, C., Ma, L.-J., Wortman, J. R., Batzoglou, S., Lee, S.-I., Baştürkmen, M., Spevak, C. C. & other authors. (2005).** Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* **438**, 1105–1115.
- Georgakopoulos, P., Lockington, R. A. & Kelly, J. M. (2012).** SAGA Complex Components and Acetate Repression in *Aspergillus nidulans*. *G3 GenesGenomesGenetics* **2**, 1357–1367.
- Geysens, S., Whyteside, G. & Archer, D. B. (2009).** Genomics of protein folding in the endoplasmic reticulum, secretion stress and glycosylation in the *Aspergilli*. *Fungal Genet Biol*, Thematic Issue: *Aspergillus* Genomics and Beyond **46**, S121–S140.
- Gielkens, M., González-Candelas, L., Sánchez-Torres, P., van de Vondervoort, P., de Graaff, L., Visser, J. & Ramón, D. (1999).** The *abfB* gene encoding the major alpha-L-arabinofuranosidase of *Aspergillus nidulans*: nucleotide sequence, regulation and construction of a disrupted strain. *Microbiol Read Engl* **145 (Pt 3)**, 735–741.
- Girard, V., Dieryckx, C., Job, C. & Job, D. (2013).** Secretomes: the fungal strike force. *Proteomics* **13**, 597–608.
- Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C. & other authors. (1996).** Life with 6000 genes. *Science* **274**, 546, 563–567.
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R. & other authors. (2011).** Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644–652.
- De Groot, M. J., Bundock, P., Hooykaas, P. J. & Beijersbergen, A. G. (1998).** *Agrobacterium tumefaciens*-mediated transformation of filamentous fungi. *Nat Biotechnol* **16**, 839–842.
- Guais, O. (2009, January 1).** *Identification des potentialités enzymatiques de Penicillium funiculosum par des approches génomiques et protéomiques*. PhD, Toulouse, INSA.

- Guais, O., Borderies, G., Pichereaux, C., Maestracci, M., Neugnot, V., Rossignol, M. & François, J. M. (2008).** Proteomics analysis of 'Rovabiot Excel', a secreted protein cocktail from the filamentous fungus *Penicillium funiculosum* grown under industrial process fermentation. *J Ind Microbiol Biotechnol* **35**, 1659–1668.
- Guais, O., Tourrassé, O., Dourdoigne, M., Parrou, J. L. & François, J. M. (2010).** Characterization of the family GH54 alpha-L-arabinofuranosidases in *Penicillium funiculosum*, including a novel protein bearing a cellulose-binding domain. *Appl Microbiol Biotechnol* **87**, 1007–1021.
- Guillemette, T., van Peij, N. N., Goosen, T., Lanthaler, K., Robson, G. D., van den Hondel, C. A., Stam, H. & Archer, D. B. (2007).** Genomic analysis of the secretion stress response in the enzyme-producing cell factory *Aspergillus niger*. *BMC Genomics* **8**, 158.
- Hacquard, S., Veneault-Fourrey, C., Delaruelle, C., Frey, P., Martin, F. & Duplessis, S. (2011).** Validation of *Melampsora larici-populina* reference genes for in planta RT-quantitative PCR expression profiling during time-course infection of poplar leaves. *Physiol Mol Plant Pathol* **75**, 106–112.
- Hacquard, S., Kracher, B., Maekawa, T., Vernaldi, S., Schulze-Lefert, P. & Ver Loren van Themaat, E. (2013).** Mosaic genome structure of the barley powdery mildew pathogen and conservation of transcriptional programs in divergent hosts. *Proc Natl Acad Sci U S A* **110**, E2219–2228.
- Hane, J. K., Anderson, J. P., Williams, A. H., Sperschneider, J. & Singh, K. B. (2014).** Genome Sequencing and Comparative Genomics of the Broad Host-Range Pathogen *Rhizoctonia solani* AG8. *PLoS Genet* **10**, e1004281.
- Harris, P. V., Welner, D., McFarland, K. C., Re, E., Navarro Poulsen, J.-C., Brown, K., Salbo, R., Ding, H., Vlasenko, E. & other authors. (2010).** Stimulation of Lignocellulosic Biomass Hydrolysis by Proteins of Glycoside Hydrolase Family 61: Structure and Function of a Large, Enigmatic Family. *Biochemistry (Mosc)* **49**, 3305–3316.

References

- Herold, S., Bischof, R., Metz, B., Seiboth, B. & Kubicek, C. P. (2013).** Xylanase gene transcription in *Trichoderma reesei* is triggered by different inducers representing different hemicellulosic pentose polymers. *Eukaryot Cell* **12**, 390–398.
- Hervás-Aguilar, A., Rodríguez, J. M., Tilburn, J., Arst, H. N., Jr & Peñalva, M. A. (2007).** Evidence for the direct involvement of the proteasome in the proteolytic processing of the *Aspergillus nidulans* zinc finger transcription factor PacC. *J Biol Chem* **282**, 34735–34747.
- Himmel, M. E., Ding, S.-Y., Johnson, D. K., Adney, W. S., Nimlos, M. R., Brady, J. W. & Foust, T. D. (2007).** Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* **315**, 804–807.
- Honda, Y., Kobayashi, K. & Kirimura, K. (2011).** Increases in gene-targeting frequencies due to disruption of *kueA* as a *ku80* homolog in citric acid-producing *Aspergillus niger*. *Biosci Biotechnol Biochem* **75**, 1594–1596.
- Houbraken, J. & Samson, R. A. (2011).** Phylogeny of *Penicillium* and the segregation of Trichocomaceae into three families. *Stud Mycol* **70**, 1–51.
- Houbraken, J., de Vries, R. P. & Samson, R. A. (2014).** Modern taxonomy of biotechnologically important *Aspergillus* and *Penicillium* species. *Adv Appl Microbiol* **86**, 199–249.
- Hruz, T., Wyss, M., Docquier, M., Pfaffl, M. W., Masanetz, S., Borghi, L., Verbrugghe, P., Kalaydjieva, L., Bleuler, S. & other authors. (2011).** RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization. *BMC Genomics* **12**, 156.
- Huang Xueling, N. A. and F. U., Feng Hao, N. A. and F. U. & Kang Zhensheng, N. A. and F. U. (2012).** Selection of reference genes for quantitative Real-time PCR normalization in *Puccinia striiformis* f.sp. *tritici*. *J Agric Biotechnol*.
- Huggett, J., Dheda, K., Bustin, S. & Zumla, A. (2005).** Real-time RT-PCR normalisation; strategies and considerations. *Genes Immun* **6**, 279–284.
- Huggett, J. & Bustin, S. A. (2011).** Standardisation and reporting for nucleic acid quantification. *Accreditation Qual Assur* **16**, 399–405.

- Hunter, A. J., Morris, T. A., Jin, B., Saint, C. P. & Kelly, J. M. (2013). Deletion of *creB* in *Aspergillus oryzae* increases secreted hydrolytic enzyme activity. *Appl Environ Microbiol* **79**, 5480–5487.
- Ichinose, S., Tanaka, M., Shintani, T. & Gomi, K. (2014). Improved α -amylase production by *Aspergillus oryzae* after a double deletion of genes involved in carbon catabolite repression. *Appl Microbiol Biotechnol* **98**, 335–343.
- Ilmén, M., Thrane, C. & Penttilä, M. (1996). The glucose repressor gene *cre1* of *Trichoderma*: isolation and expression of a full-length and a truncated mutant form. *Mol Gen Genet MGG* **251**, 451–460.
- Jiang, D., Zhu, W., Wang, Y., Sun, C., Zhang, K.-Q. & Yang, J. (2013). Molecular tools for functional genomics in filamentous fungi: Recent advances and new strategies. *Biotechnol Adv* **31**, 1562–1574.
- Jørgensen, T. R., Goosen, T., Hondel, C. A. van den, Ram, A. F. & Iversen, J. J. (2009). Transcriptomic comparison of *Aspergillus niger* growing on two different sugars reveals coordinated regulation of the secretory pathway. *BMC Genomics* **10**, 1–16.
- Kavanagh, K. (2011). *Fungi: Biology and Applications*. John Wiley & Sons.
- Kim, H. K. & Yun, S. H. (2011). Evaluation of Potential Reference Genes for Quantitative RT-PCR Analysis in *Fusarium graminearum* under Different Culture Conditions. *Plant Pathol J*.
- Kozera, B. & Rapacz, M. (2013). Reference genes in real-time PCR. *J Appl Genet* **54**, 391–406.
- Kramer, C., Loros, J. J., Dunlap, J. C. & Crosthwaite, S. K. (2003). Role for antisense RNA in regulating circadian clock function in *Neurospora crassa*. *Nature* **421**, 948–952.
- Krappmann, S., Sasse, C. & Braus, G. H. (2006). Gene Targeting in *Aspergillus fumigatus* by Homologous Recombination Is Facilitated in a Nonhomologous End- Joining-Deficient Genetic Background. *Eukaryot Cell* **5**, 212–215.
- Krijgheld, P., Bleichrodt, R., van Veluw, G. J., Wang, F., Müller, W. H., Dijksterhuis, J. & Wösten, H. A. B. (2013). Development in *Aspergillus*. *Stud Mycol* **74**, 1–29.

References

- Kück, U. & Hoff, B. (2010).** New tools for the genetic manipulation of filamentous fungi. *Appl Microbiol Biotechnol* **86**, 51–62.
- Kurasawa, T., Yachi, M., Suto, M., Kamagata, Y., Takao, S. & Tomita, F. (1992).** Induction of Cellulase by Gentiobiose and Its Sulfur-Containing Analog in *Penicillium purpurogenum*. *Appl Environ Microbiol* **58**, 106–110.
- De La Mare, M. (2014).** *Caractérisation des x-L-arabinofuranosidases de la famille GH62 chez Talaromyces versatilis et étude de leur impact, en association avec des xylanases, sur la dégradation d'arabinoxylane*. PhD, Toulouse, INSA.
- Lafond, M., Guais, O., Maestracci, M., Bonnin, E. & Giardina, T. (2014).** Four GH11 xylanases from the xylanolytic fungus *Talaromyces versatilis* act differently on (arabino)xylans. *Appl Microbiol Biotechnol*.
- Lagaert, S., Beliën, T. & Volckaert, G. (2009).** Plant cell walls: Protecting the barrier from degradation by microbial enzymes. *Semin Cell Dev Biol* **20**, 1064–1073.
- Lamarre, C., Sokol, S., Debeaupuis, J.-P., Henry, C., Lacroix, C., Glaser, P., Coppée, J.-Y., François, J.-M. & Latgé, J.-P. (2008).** Transcriptomic analysis of the exit from dormancy of *Aspergillus fumigatus* conidia. *BMC Genomics* **9**, 417.
- Langston, J. A., Shaghasi, T., Abbate, E., Xu, F., Vlasenko, E. & Sweeney, M. D. (2011).** Oxidoreductive cellulose depolymerization by the enzymes cellobiose dehydrogenase and glycoside hydrolase 61. *Appl Environ Microbiol* **77**, 7007–7015.
- Latgé, J.-P. & Beauvais, A. (2014).** Functional duality of the cell wall. *Curr Opin Microbiol, Host-microbe interactions: fungi/parasites/viruses* **20**, 111–117.
- Laurell, H., Iacovoni, J. S., Abot, A., Svec, D., Maoret, J.-J., Arnal, J.-F. & Kubista, M. (2012).** Correction of RT-qPCR data for genomic DNA-derived signals with ValidPrime. *Nucleic Acids Res* **40**, e51.

- Leeuwen, M. R. van, Krijgheld, P., Bleichrodt, R., Menke, H., Stam, H., Stark, J., Wösten, H. a. B. & Dijksterhuis, J. (2013). Germination of conidia of *Aspergillus niger* is accompanied by major changes in RNA profiles. *Stud Mycol* **74**, 59–70.
- Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M. & Henrissat, B. (2013). Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnol Biofuels* **6**, 41.
- Li, L., Chang, S. & Liu, Y. (2010). RNA interference pathways in filamentous fungi. *Cell Mol Life Sci CMLS* **67**, 3849–3863.
- Lin, J.-Q., Zhao, X.-X., Zhi, Q.-Q., Zhao, M. & He, Z.-M. (2013). Transcriptomic profiling of *Aspergillus flavus* in response to 5-azacytidine. *Fungal Genet Biol FG B* **56**, 78–86.
- Linder, M. B., Szilvay, G. R., Nakari-Setälä, T. & Penttilä, M. E. (2005). Hydrophobins: the protein-amphiphiles of filamentous fungi. *FEMS Microbiol Rev* **29**, 877–896.
- Liu, Z. L. & Slininger, P. J. (2007). Universal external RNA controls for microbial gene expression analysis using microarray and qRT-PCR. *J Microbiol Methods* **68**, 486–496.
- Lockington, R. A. & Kelly, J. M. (2001). Carbon catabolite repression in *Aspergillus nidulans* involves deubiquitination. *Mol Microbiol* **40**, 1311–1321.
- Lockington, R. A. & Kelly, J. M. (2002). The WD40-repeat protein CreC interacts with and stabilizes the deubiquitinating enzyme CreB in vivo in *Aspergillus nidulans*. *Mol Microbiol* **43**, 1173–1182.
- Lockington, R. A., Rodbourn, L., Barnett, S., Carter, C. J. & Kelly, J. M. (2002). Regulation by carbon and nitrogen sources of a family of cellulases in *Aspergillus nidulans*. *Fungal Genet Biol FG B* **37**, 190–196.
- Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **42**, D490–495.

References

- Lozano-Ojalvo, D., Rodríguez, A., Bernáldez, V., Córdoba, J. J. & Rodríguez, M. (2013).** Influence of temperature and substrate conditions on the omt-1 gene expression of *Aspergillus parasiticus* in relation to its aflatoxin production. *Int J Food Microbiol* **166**, 263–269.
- MacCabe, A. P., Orejas, M., Pérez-González, J. A. & Ramón, D. (1998).** Opposite patterns of expression of two *Aspergillus nidulans* xylanase genes with respect to ambient pH. *J Bacteriol* **180**, 1331–1333.
- Mach-Aigner, A. R., Pucher, M. E., Steiger, M. G., Bauer, G. E., Preis, S. J. & Mach, R. L. (2008).** Transcriptional regulation of *xyr1*, encoding the main regulator of the xylanolytic and cellulolytic enzyme system in *Hypocrea jecorina*. *Appl Environ Microbiol* **74**, 6554–6562.
- Mach-Aigner, A. R., Omony, J., Jovanovic, B., van Boxel, A. J. B. & de Graaff, L. H. (2012).** d-Xylose concentration-dependent hydrolase expression profiles and the function of CreA and XlnR in *Aspergillus niger*. *Appl Environ Microbiol* **78**, 3145–3155.
- Machida, M., Asai, K., Sano, M., Tanaka, T., Kumagai, T., Terai, G., Kusumoto, K.-I., Arima, T., Akita, O. & other authors. (2005).** Genome sequencing and analysis of *Aspergillus oryzae*. *Nature* **438**, 1157–1161.
- MacPherson, S., Larochelle, M. & Turcotte, B. (2006).** A Fungal Family of Transcriptional Regulators: the Zinc Cluster Proteins. *Microbiol Mol Biol Rev* **70**, 583–604.
- Mannhaupt, G., Montrone, C., Haase, D., Mewes, H. W., Aign, V., Hoheisel, J. D., Fartmann, B., Nyakatura, G., Kempken, F. & other authors. (2003).** What's in the genome of a filamentous fungus? Analysis of the *Neurospora* genome sequence. *Nucleic Acids Res* **31**, 1944–1954.
- Margeot, A., Hahn-Hagerdal, B., Edlund, M., Slade, R. & Monot, F. (2009).** New improvements for lignocellulosic ethanol. *Curr Opin Biotechnol* **20**, 372–380.
- Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. (2008).** RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**, 1509–1517.

- Martinez, D., Larrondo, L. F., Putnam, N., Gelpke, M. D. S., Huang, K., Chapman, J., Helfenbein, K. G., Ramaiya, P., Detter, J. C. & other authors. (2004).** Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nat Biotechnol* **22**, 695–700.
- Martinez, D., Berka, R. M., Henrissat, B., Saloheimo, M., Arvas, M., Baker, S. E., Chapman, J., Chertkov, O., Coutinho, P. M. & other authors. (2008).** Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). *Nat Biotechnol* **26**, 553–560.
- Marui, J., Tanaka, A., Mimura, S., de Graaff, L. H., Visser, J., Kitamoto, N., Kato, M., Kobayashi, T. & Tsukagoshi, N. (2002).** A Transcriptional Activator, AoXlnR, Controls the Expression of Genes Encoding Xylanolytic Enzymes in *Aspergillus oryzae*. *Fungal Genet Biol* **35**, 157–169.
- Mathieu, M. & Felenbok, B. (1994).** The *Aspergillus nidulans* CREA protein mediates glucose repression of the ethanol regulon at various levels through competition with the ALCR-specific transactivator. *EMBO J* **13**, 4022–4027.
- Mathieu, M., Nikolaev, I., Scazzocchio, C. & Felenbok, B. (2005).** Patterns of nucleosomal organization in the alc regulon of *Aspergillus nidulans*: roles of the AlcR transcriptional activator and the CreA global repressor. *Mol Microbiol* **56**, 535–548.
- Maxam, A. M. & Gilbert, W. (1977).** A new method for sequencing DNA. *Proc Natl Acad Sci* **74**, 560–564.
- McKelvey, S. M. & Murphy, R. A. (2010).** Analysis of wide-domain transcriptional regulation in solid-state cultures of *Aspergillus oryzae*. *J Ind Microbiol Biotechnol* **37**, 455–469.
- Metzker, M. L. (2010).** Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31–46.
- Meyer, V. (2008).** Genetic engineering of filamentous fungi - progress, obstacles and future trends. *Biotechnol Adv* **26**, 177–185.

References

- Meyer, V., Arentshorst, M., El-Ghezal, A., Drews, A.-C., Kooistra, R., van den Hondel, C. A. M. J. J. & Ram, A. F. J. (2007).** Highly efficient gene targeting in the *Aspergillus niger kusA* mutant. *J Biotechnol* **128**, 770–775.
- Michielse, C. B., Hooykaas, P. J. J., van den Hondel, C. A. M. J. J. & Ram, A. F. J. (2005).** *Agrobacterium*-mediated transformation as a tool for functional genomics in fungi. *Curr Genet* **48**, 1–17.
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. (2008).** Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621–628.
- Mulder, H. J., Saloheimo, M., Penttilä, M. & Madrid, S. M. (2004).** The transcription factor HACA mediates the unfolded protein response in *Aspergillus niger*, and up-regulates its own transcription. *Mol Genet Genomics* **271**, 130–140.
- Mulder, H. J., Nikolaev, I. & Madrid, S. M. (2006).** HACA, the transcriptional activator of the unfolded protein response (UPR) in *Aspergillus niger*, binds to partly palindromic UPR elements of the consensus sequence 5'-CAN(G/A)NTGT/GCCT-3'. *Fungal Genet Biol FG B* **43**, 560–572.
- Mulder, H. J. & Nikolaev, I. (2009).** HacA-Dependent Transcriptional Switch Releases hacA mRNA from a Translational Block upon Endoplasmic Reticulum Stress. *Eukaryot Cell* **8**, 665–675.
- Munro, C. A. (2013).** Chapter Four - Chitin and Glucan, the Yin and Yang of the Fungal Cell Wall, Implications for Antifungal Drug Discovery and Therapy. In *Adv Appl Microbiol*, pp. 145–172. Edited by Sima Sariaslani and Geoffrey M. Gadd. Academic Press.
- Navarrete, K., Roa, A., Vaca, I., Espinosa, Y., Navarro, C. & Chávez, R. (2009).** Molecular characterization of the *niaD* and *pyrG* genes from *Penicillium camemberti*, and their use as transformation markers. *Cell Mol Biol Lett* **14**, 692–702.
- Nayak, T., Szewczyk, E., Oakley, C. E., Osmani, A., Ukil, L., Murray, S. L., Hynes, M. J., Osmani, S. A. & Oakley, B. R. (2006).** A Versatile and Efficient Gene-Targeting System for *Aspergillus nidulans*. *Genetics* **172**, 1557–1566.

- Nielsen, J. B., Nielsen, M. L. & Mortensen, U. H. (2008).** Transient disruption of non-homologous end-joining facilitates targeted genome manipulations in the filamentous fungus *Aspergillus nidulans*. *Fungal Genet Biol FG B* **45**, 165–170.
- Nielsen, M. L., Albertsen, L., Lettier, G., Nielsen, J. B. & Mortensen, U. H. (2006).** Efficient PCR-based gene targeting with a recyclable marker for *Aspergillus nidulans*. *Fungal Genet Biol FG B* **43**, 54–64.
- Nierman, W. C., Pain, A., Anderson, M. J., Wortman, J. R., Kim, H. S., Arroyo, J., Berriman, M., Abe, K., Archer, D. B. & other authors. (2005).** Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature* **438**, 1151–1156.
- Ninomiya, Y., Suzuki, K., Ishii, C. & Inoue, H. (2004).** Highly efficient gene replacements in *Neurospora* strains deficient for nonhomologous end-joining. *Proc Natl Acad Sci U S A* **101**, 12248–12253.
- Nitsche, B. M., Jørgensen, T. R., Akeroyd, M., Meyer, V. & Ram, A. F. (2012).** The carbon starvation response of *Aspergillus niger* during submerged cultivation: Insights from the transcriptome and secretome. *BMC Genomics* **13**, 380.
- Noguchi, Y., Sano, M., Kanamaru, K., Ko, T., Takeuchi, M., Kato, M. & Kobayashi, T. (2009).** Genes regulated by AoXlnR, the xylanolytic and cellulolytic transcriptional regulator, in *Aspergillus oryzae*. *Appl Microbiol Biotechnol* **85**, 141–154.
- Noguchi, Y., Tanaka, H., Kanamaru, K., Kato, M. & Kobayashi, T. (2011).** Xylose triggers reversible phosphorylation of XlnR, the fungal transcriptional activator of xylanolytic and cellulolytic genes in *Aspergillus oryzae*. *Biosci Biotechnol Biochem* **75**, 953–959.
- Novodvorska, M., Hayer, K., Pullan, S. T., Wilson, R., Blythe, M. J., Stam, H., Stratford, M. & Archer, D. B. (2013).** Transcriptional landscape of *Aspergillus niger* at breaking of conidial dormancy revealed by RNA-sequencing. *BMC Genomics* **14**, 246.
- Nowrousian, M., Stajich, J. E., Chu, M., Engh, I., Espagne, E., Halliday, K., Kamerewerd, J., Kempken, F., Knab, B. & other authors. (2010).** De novo assembly of a 40 Mb eukaryotic genome from

References

- short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. *PLoS Genet* **6**, e1000891.
- O'Connell, R. J., Thon, M. R., Hacquard, S., Amyotte, S. G., Kleemann, J., Torres, M. F., Damm, U., Buiate, E. A., Epstein, L. & other authors. (2012).** Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses. *Nat Genet* **44**, 1060–1065.
- Ogawa, M., Kobayashi, T. & Koyama, Y. (2012a).** ManR, a transcriptional regulator of the β -mannan utilization system, controls the cellulose utilization system in *Aspergillus oryzae*. *Biosci Biotechnol Biochem* **77**, 426–429.
- Ogawa, M., Kobayashi, T. & Koyama, Y. (2012b).** ManR, a novel Zn(II)₂Cys₆ transcriptional activator, controls the β -mannan utilization system in *Aspergillus oryzae*. *Fungal Genet Biol* **49**, 987–995.
- Ohtaki, S., Maeda, H., Takahashi, T., Yamagata, Y., Hasegawa, F., Gomi, K., Nakajima, T. & Abe, K. (2006).** Novel hydrophobic surface binding protein, HsbA, produced by *Aspergillus oryzae*. *Appl Environ Microbiol* **72**, 2407–2413.
- Osmanova, N., Schultze, W. & Ayoub, N. (2010).** Azaphilones: a class of fungal metabolites with diverse biological activities. *Phytochem Rev* **9**, 315–342.
- Pakula, T. M., Laxell, M., Huuskonen, A., Uusitalo, J., Saloheimo, M. & Penttilä, M. (2003).** The effects of drugs inhibiting protein secretion in the filamentous fungus *Trichoderma reesei*. Evidence for down-regulation of genes that encode secreted proteins in the stressed cells. *J Biol Chem* **278**, 45011–45020.
- Park, E. & Rapoport, T. A. (2012).** Mechanisms of Sec61/SecY-mediated protein translocation across membranes. *Annu Rev Biophys* **41**, 21–40.
- Van Peij, N. N., Visser, J. & de Graaff, L. H. (1998a).** Isolation and analysis of *xlnR*, encoding a transcriptional activator co-ordinating xylanolytic expression in *Aspergillus niger*. *Mol Microbiol* **27**, 131–142.

- Van Peij, N. N. M. E., Gielkens, M. M. C., de Vries, R. P., Visser, J. & de Graaff, L. H. (1998b).** The Transcriptional Activator XlnR Regulates Both Xylanolytic and Endoglucanase Gene Expression in *Aspergillus niger*. *Appl Environ Microbiol* **64**, 3615–3619.
- Pel, H. J., de Winde, J. H., Archer, D. B., Dyer, P. S., Hofmann, G., Schaap, P. J., Turner, G., de Vries, R. P., Albang, R. & other authors. (2007).** Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. *Nat Biotechnol* **25**, 221–231.
- Peñalva, M. A., Tilburn, J., Bignell, E. & Arst, H. N., Jr. (2008).** Ambient pH gene regulation in fungi: making connections. *Trends Microbiol* **16**, 291–300.
- Penas, M. M., Hervas-Aguilar, A., Munera-Huertas, T., Reoyo, E., Penalva, M. A., Arst, H. N. & Tilburn, J. (2007).** Further Characterization of the Signaling Proteolysis Step in the *Aspergillus nidulans* pH Signal Transduction Pathway. *Eukaryot Cell* **6**, 960–970.
- Perez-Gonzalez, J. A., van Peij, N. N. M. E., Bezoen, A., Maccabe, A. P., Ramon, D. & de Graaff, L. H. (1998).** Molecular Cloning and Transcriptional Regulation of the *Aspergillus nidulans* *xlnD* Gene Encoding beta-Xylosidase. *Appl Environ Microbiol* **64**, 1412–1419.
- Pfaffl, M. W., Tichopad, A., Prgomet, C. & Neuvians, T. P. (2004).** Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper – Excel-based tool using pair-wise correlations. *Biotechnol Lett* **26**, 509–515.
- Ponting, C. P., Oliver, P. L. & Reik, W. (2009).** Evolution and Functions of Long Noncoding RNAs. *Cell* **136**, 629–641.
- Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. & Gu, Y. (2012).** A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341.
- Raffaello, T. & Asiegbu, F. O. (2013).** Evaluation of potential reference genes for use in gene expression studies in the conifer pathogen (*Heterobasidion annosum*). *Mol Biol Rep* **40**, 4605–4611.

References

- Raggam, R. B., Salzer, H. J. F., Marth, E., Heiling, B., Paulitsch, A. H. & Buzina, W. (2011).** Molecular detection and characterisation of fungal heat shock protein 60. *Mycoses* **54**, e394–e399.
- Ramsden, S. C., Daly, S., Geilenkeuser, W.-J., Duncan, G., Hermitte, F., Marubini, E., Neumaier, M., Orlando, C., Palicka, V. & other authors. (2006).** EQUAL-quant: an international external quality assessment scheme for real-time PCR. *Clin Chem* **52**, 1584–1591.
- Rauscher, R., Würleitner, E., Wacenovsky, C., Aro, N., Stricker, A. R., Zeilinger, S., Kubicek, C. P., Penttilä, M. & Mach, R. L. (2006).** Transcriptional regulation of *xyn1*, encoding xylanase I, in *Hypocrea jecorina*. *Eukaryot Cell* **5**, 447–456.
- Reinhard Fischer. (2002).** Conidiation in *Aspergillus nidulans*. In *Mol Biol Fungal Dev*. CRC Press.
- Ries, L., Pullan, S. T., Delmas, S., Malla, S., Blythe, M. J. & Archer, D. B. (2013).** Genome-wide transcriptional response of *Trichoderma reesei* to lignocellulose using RNA sequencing and comparison with *Aspergillus niger*. *BMC Genomics* **14**, 541.
- Rivera, A. L., Magaña-Ortíz, D., Gómez-Lim, M., Fernández, F. & Loske, A. M. (2014).** Physical methods for genetic transformation of fungi and yeast. *Phys Life Rev* **11**, 184–203.
- Rolland, S., Jobic, C., Fèvre, M. & Bruel, C. (2003).** *Agrobacterium*-mediated transformation of *Botrytis cinerea*, simple purification of monokaryotic transformants and rapid conidia-based identification of the transfer-DNA host genomic DNA flanking sequences. *Curr Genet* **44**, 164–171.
- Rouxel, T., Grandaubert, J., Hane, J. K., Hoede, C., van de Wouw, A. P., Couloux, A., Dominguez, V., Anthouard, V., Bally, P. & other authors. (2011).** Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. *Nat Commun* **2**, 202.
- Ruiz-Díez, B. (2002).** Strategies for the transformation of filamentous fungi. *J Appl Microbiol* **92**, 189–195.

- Saloheimo, A., Aro, N., Ilmén, M. & Penttilä, M. (2000).** Isolation of the *ace1* gene encoding a Cys(2)-His(2) transcription factor involved in regulation of activity of the cellulase promoter *cbh1* of *Trichoderma reesei*. *J Biol Chem* **275**, 5817–5825.
- Saloheimo, M. & Pakula, T. M. (2012).** The cargo and the transport system: secreted proteins and protein secretion in *Trichoderma reesei* (*Hypocrea jecorina*). *Microbiol Read Engl* **158**, 46–57.
- Saloheimo, M., Paloheimo, M., Hakola, S., Pere, J., Swanson, B., Nyysönen, E., Bhatia, A., Ward, M. & Penttilä, M. (2002).** Swollenin, a *Trichoderma reesei* protein with sequence similarity to the plant expansins, exhibits disruption activity on cellulosic materials. *Eur J Biochem FEBS* **269**, 4202–4211.
- Saloheimo, M., Valkonen, M. & Penttilä, M. (2003).** Activation mechanisms of the HAC1-mediated unfolded protein response in filamentous fungi. *Mol Microbiol* **47**, 1149–1161.
- Sambrook, J. & Russell, D. W. (2001).** *Molecular Cloning: A Laboratory Manual*. CSHL Press.
- Samson, R. A., Yilmaz, N., Houbraken, J., Spierenburg, H., Seifert, K. A., Peterson, S. W., Varga, J. & Frisvad, J. C. (2011).** Phylogeny and nomenclature of the genus *Talaromyces* and taxa accommodated in *Penicillium* subgenus *Biverticillium*. *Stud Mycol* **70**, 159–183.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977).** DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–5467.
- Santangelo, G. M. (2006).** Glucose signaling in *Saccharomyces cerevisiae*. *Microbiol Mol Biol Rev MMBR* **70**, 253–282.
- Schmidt, S. M. & Panstruga, R. (2011).** Pathogenomics of fungal plant parasites: what have we learnt about pathogenesis? *Curr Opin Plant Biol* **14**, 392–399.
- Seidl, V., Gamauf, C., Druzhinina, I. S., Seiboth, B., Hartl, L. & Kubicek, C. P. (2008).** The *Hypocrea jecorina* (*Trichoderma reesei*) hypercellulolytic mutant RUT C30 lacks a 85 kb (29 gene-encoding) region of the wild-type genome. *BMC Genomics* **9**, 327.
- Shroff, R. A., Lockington, R. A. & Kelly, J. M. (1996).** Analysis of mutations in the *creA* gene involved in carbon catabolite repression in *Aspergillus nidulans*. *Can J Microbiol* **42**, 950–959.

References

- Shroff, R. A., O'Connor, S. M., Hynes, M. J., Lockington, R. A. & Kelly, J. M. (1997).** Null alleles of *creA*, the regulator of carbon catabolite repression in *Aspergillus nidulans*. *Fungal Genet Biol FG B* **22**, 28–38.
- Sibthorp, C., Wu, H., Cowley, G., Wong, P. W. H., Palaima, P., Morozov, I. Y., Weedall, G. D. & Caddick, M. X. (2013).** Transcriptome analysis of the filamentous fungus *Aspergillus nidulans* directed to the global identification of promoters. *BMC Genomics* **14**, 847.
- Sigoillot, J.-C., Berrin, J.-G., Bey, M., Lesage-Meessen, L., Levasseur, A., Lomascolo, A., Record, E. & Uzan-Boukhris, E. (2012).** Chapter 8 - Fungal Strategies for Lignin Degradation. In *Adv Bot Res*, Lignins Biosynthesis, Biodegradation and Bioengineering, pp. 263–308. Edited by Lise Jouanin and Catherine Lapierre. Academic Press.
- Da Silva Ferreira, M. E., Kress, M. R. V. Z., Savoldi, M., Goldman, M. H. S., Härtl, A., Heinekamp, T., Brakhage, A. A. & Goldman, G. H. (2006).** The *akuB* (KU80) mutant deficient for nonhomologous end joining is a powerful tool for analyzing pathogenicity in *Aspergillus fumigatus*. *Eukaryot Cell* **5**, 207–211.
- Silver, N., Best, S., Jiang, J. & Thein, S. L. (2006).** Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol Biol* **7**, 33.
- Smith, L. G. (2001).** Plant cell division: building walls in the right places. *Nat Rev Mol Cell Biol* **2**, 33–39.
- Soanes, D. M., Chakrabarti, A., Paszkiewicz, K. H., Dawe, A. L. & Talbot, N. J. (2012).** Genome-wide transcriptional profiling of appressorium development by the rice blast fungus *Magnaporthe oryzae*. *PLoS Pathog* **8**, e1002514.
- De Souza, W. R., Maitan-Alfenas, G. P., de Gouvêa, P. F., Brown, N. A., Savoldi, M., Battaglia, E., Goldman, M. H. S., de Vries, R. P. & Goldman, G. H. (2013).** The influence of *Aspergillus niger* transcription factors AraR and XlnR in the gene expression during growth in D-xylose, L-arabinose and steam-exploded sugarcane bagasse. *Fungal Genet Biol FG B* **60**, 29–45.
- Spang, A. (2008).** The life cycle of a transport vesicle. *Cell Mol Life Sci CMLS* **65**, 2781–2789.

- Spanu, P. D., Abbott, J. C., Amselem, J., Burgis, T. A., Soanes, D. M., Stüber, K., Ver Loren van Themaat, E., Brown, J. K. M., Butcher, S. A. & other authors. (2010). Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* **330**, 1543–1546.
- Steen, J. A. & Cooper, M. A. (2011). Fluorogenic pyrosequencing in microreactors. *Nat Methods* **8**, 548–549.
- Steiger, M. G., Mach, R. L. & Mach-Aigner, A. R. (2010). An accurate normalization strategy for RT-qPCR in *Hypocrea jecorina* (*Trichoderma reesei*). *J Biotechnol* **145**, 30–37.
- Steinberg, G. (2007). Hyphal Growth: a Tale of Motors, Lipids, and the Spitzenkörper. *Eukaryot Cell* **6**, 351–360.
- Steyaert, J. M., Weld, R. J., Mendoza-Mendoza, A. & Stewart, A. (2010). Reproduction without sex: conidiation in the filamentous fungus *Trichoderma*. *Microbiol Read Engl* **156**, 2887–2900.
- Stranneheim, H. & Lundeberg, J. (2012). Stepping stones in DNA sequencing. *Biotechnol J* **7**, 1063–1073.
- Stricker, A. R., Grosstessner-Hain, K., Würleitner, E. & Mach, R. L. (2006). Xyr1 (xylanase regulator 1) regulates both the hydrolytic enzyme system and D-xylose metabolism in *Hypocrea jecorina*. *Eukaryot Cell* **5**, 2128–2137.
- Stricker, A. R., Trefflinger, P., Aro, N., Penttilä, M. & Mach, R. L. (2008). Role of Ace2 (Activator of Cellulases 2) within the *xyn2* transcriptosome of *Hypocrea jecorina*. *Fungal Genet Biol FG B* **45**, 436–445.
- Stukenbrock, E. H. & Croll, D. (2014). The evolving fungal genome. *Fungal Biol Rev* **28**, 1–12.
- Sun, J. & Glass, N. L. (2011). Identification of the CRE-1 Cellulolytic Regulon in *Neurospora crassa*. *PLoS ONE* **6**, e25654.
- Sun, J., Tian, C., Diamond, S. & Glass, N. L. (2012). Deciphering transcriptional regulatory mechanisms associated with hemicellulose degradation in *Neurospora crassa*. *Eukaryot Cell* **11**, 482–493.

References

- Takahashi, T., Masuda, T. & Koyama, Y. (2006).** Enhanced gene targeting frequency in *ku70* and *ku80* disruption mutants of *Aspergillus sojae* and *Aspergillus oryzae*. *Mol Genet Genomics* **MGG 275**, 460–470.
- Takahashi, T., Maeda, H., Yoneda, S., Ohtaki, S., Yamagata, Y., Hasegawa, F., Gomi, K., Nakajima, T. & Abe, K. (2005).** The fungal hydrophobin RolA recruits polyesterase and laterally moves on hydrophobic surfaces. *Mol Microbiol* **57**, 1780–1796.
- Tamayo, E. N., Villanueva, A., Hasper, A. A., de Graaff, L. H., Ramón, D. & Orejas, M. (2008).** CreA mediates repression of the regulatory gene *xlnR* which controls the production of xylanolytic enzymes in *Aspergillus nidulans*. *Fungal Genet Biol FG B* **45**, 984–993.
- Tang, J. D., Parker, L. A., Perkins, A. D., Sonstegard, T. S., Schroeder, S. G., Nicholas, D. D. & Diehl, S. V. (2013).** Gene expression analysis of copper tolerance and wood decay in the brown rot fungus *Fibroporia radiculosa*. *Appl Environ Microbiol* **79**, 1523–1533.
- Teichert, I., Wolff, G., Kück, U. & Nowrousian, M. (2012).** Combining laser microdissection and RNA-seq to chart the transcriptional landscape of fungal development. *BMC Genomics* **13**, 511.
- Teste, M.-A., Duquenne, M., François, J. M. & Parrou, J.-L. (2009).** Validation of reference genes for quantitative expression analysis by real-time RT-PCR in *Saccharomyces cerevisiae*. *BMC Mol Biol* **10**, 99.
- Thellin, O., ElMoualij, B., Heinen, E. & Zorzi, W. (2009).** A decade of improvements in quantification of gene expression and internal standard selection. *Biotechnol Adv* **27**, 323–333.
- Thornton, C. R. & Wills, O. E. (2013).** Immunodetection of fungal and oomycete pathogens: Established and emerging threats to human health, animal welfare and global food security. *Crit Rev Microbiol* 1–25.
- Tian, C., Beeson, W. T., Iavarone, A. T., Sun, J., Marletta, M. A., Cate, J. H. D. & Glass, N. L. (2009).** Systems analysis of plant cell wall degradation by the model filamentous fungus *Neurospora crassa*. *Proc Natl Acad Sci U S A* **106**, 22157–22162.
- Tijsterman, M. & Plasterk, R. H. A. (2004).** Dicers at RISC; the mechanism of RNAi. *Cell* **117**, 1–3.

- Tisserant, E., Malbreil, M., Kuo, A., Kohler, A., Symeonidi, A., Balestrini, R., Charron, P., Duensing, N., Frey, N. F. dit & other authors. (2013). Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proc Natl Acad Sci* 201313452.
- Todd, R. B., Lockington, R. A. & Kelly, J. M. (2000). The *Aspergillus nidulans creC* gene involved in carbon catabolite repression encodes a WD40 repeat protein. *Mol Gen Genet MGG* **263**, 561–570.
- Traeger, S., Altegoer, F., Freitag, M., Gabaldon, T., Kempken, F., Kumar, A., Marcet-Houben, M., Pöggeler, S., Stajich, J. E. & Nowrousian, M. (2013). The genome and development-dependent transcriptomes of *Pyronema confluens*: a window into fungal evolution. *PLoS Genet* **9**, e1003820.
- Twumasi-Boateng, K., Yu, Y., Chen, D., Gravelat, F. N., Nierman, W. C. & Sheppard, D. C. (2009). Transcriptional profiling identifies a role for BrlA in the response to nitrogen depletion and for StuA in the regulation of secondary metabolite clusters in *Aspergillus fumigatus*. *Eukaryot Cell* **8**, 104–115.
- Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A. & Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* **3**, RESEARCH0034.
- Vandesompele, J., Kubista, M. & Pfaffl, M. W. (2009). Reference gene validation software for improved normalization. In *Real-Time PCR Curr Technol Appl*, Logan J, Edwards K, Saunders N., pp. 47–64. Caister Academic Press.
- Vermeulen, J., De Preter, K., Lefever, S., Nuytens, J., De Vloed, F., Derveaux, S., Hellemans, J., Speleman, F. & Vandesompele, J. (2011). Measurable impact of RNA quality on gene expression results from quantitative PCR. *Nucleic Acids Res* **39**, e63.
- Vieira, A., Talhinhos, P., Loureiro, A., Duplessis, S., Fernandez, D., Silva, M. do C., Paulo, O. S. & Azinheira, H. G. (2011). Validation of RT-qPCR reference genes for in planta expression studies in *Hemileia vastatrix*, the causal agent of coffee leaf rust. *Fungal Biol* **115**, 891–901.

References

- De Vries, R. P. (2003).** Regulation of *Aspergillus* genes encoding plant cell wall polysaccharide-degrading enzymes; relevance for industrial production. *Appl Microbiol Biotechnol* **61**, 10–20.
- De Vries, R. P. & Visser, J. (2001).** *Aspergillus* Enzymes Involved in Degradation of Plant Cell Wall Polysaccharides. *Microbiol Mol Biol Rev* **65**, 497–522.
- Wang, B., Guo, G., Wang, C., Lin, Y., Wang, X., Zhao, M., Guo, Y., He, M., Zhang, Y. & Pan, L. (2010a).** Survey of the transcriptome of *Aspergillus oryzae* via massively parallel mRNA sequencing. *Nucleic Acids Res* **38**, 5075–5087.
- Wang, L., Feng, Z., Wang, X., Wang, X. & Zhang, X. (2010b).** DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**, 136–138.
- Wang, Y., Tang, R., Tao, J., Gao, G., Wang, X., Mu, Y. & Feng, Y. (2011).** Quantitative investigation of non-hydrolytic disruptive activity on crystalline cellulose and application to recombinant swollenin. *Appl Microbiol Biotechnol* **91**, 1353–1363.
- Wiemann, P. & Keller, N. P. (2014).** Strategies for mining fungal natural products. *J Ind Microbiol Biotechnol* **41**, 301–313.
- Wiemann, P., Sieber, C. M. K., von Bargen, K. W., Studt, L., Niehaus, E.-M., Espino, J. J., Huß, K., Michielse, C. B., Albermann, S. & other authors. (2013).** Deciphering the Cryptic Genome: Genome-wide Analyses of the Rice Pathogen *Fusarium fujikuroi* Reveal Complex Regulation of Secondary Metabolism and Novel Metabolites. *PLoS Pathog* **9**, e1003475.
- Wittmann, J. G., Heinrich, D., Gasow, K., Frey, A., Diederichsen, U. & Rudolph, M. G. (2008).** Structures of the human orotidine-5'-monophosphate decarboxylase support a covalent mechanism and provide a framework for drug design. *Struct Lond Engl* **16**, 82–92.
- Wong, M. L. & Medrano, J. F. (2005).** Real-time PCR for mRNA quantitation. *BioTechniques* **39**, 75–85.
- Woo, P. C. Y., Lau, S. K. P., Liu, B., Cai, J. J., Chong, K. T. K., Tse, H., Kao, R. Y. T., Chan, C.-M., Chow, W.-N. & Yuen, K.-Y. (2011).** Draft genome sequence of *Penicillium marneffeii* strain PM1. *Eukaryot Cell* **10**, 1740–1741.

- Wortman, J. R., Gilsenan, J. M., Joardar, V., Deegan, J., Clutterbuck, J., Andersen, M. R., Archer, D., Bencina, M., Braus, G. & other authors. (2009). The 2008 update of the *Aspergillus nidulans* genome annotation: a community effort. *Fungal Genet Biol FG B* **46 Suppl 1**, S2–13.
- Xu, Q., Singh, A. & Himmel, M. E. (2009). Perspectives and new directions for the production of bioethanol using consolidated bioprocessing of lignocellulose. *Curr Opin Biotechnol* **20**, 364–371.
- Yan, H.-Z. & Liou, R.-F. (2006). Selection of internal control genes for real-time quantitative RT-PCR assays in the oomycete plant pathogen *Phytophthora parasitica*. *Fungal Genet Biol FG B* **43**, 430–438.
- Yu, J., Fedorova, N. D., Montalbano, B. G., Bhatnagar, D., Cleveland, T. E., Bennett, J. W. & Nierman, W. C. (2011). Tight control of mycotoxin biosynthesis gene expression in *Aspergillus flavus* by temperature as revealed by RNA-Seq. *FEMS Microbiol Lett* **322**, 145–149.
- Yuen, K.-Y., Pascal, G., Wong, S. S. Y., Glaser, P., Woo, P. C. Y., Kunst, F., Cai, J. J., Cheung, E. Y. L., Médigue, C. & Danchin, A. (2003). Exploring the *Penicillium marneffeii* genome. *Arch Microbiol* **179**, 339–353.
- Zampieri, D., Nora, L. C., Basso, V., Camassola, M. & Dillon, A. J. P. (2014). Validation of reference genes in *Penicillium echinulatum* to enable gene expression study using real-time quantitative RT-PCR. *Curr Genet*.
- Zhang, J., Mao, Z., Xue, W., Li, Y., Tang, G., Wang, A., Zhang, Y. & Wang, H. (2011a). Ku80 gene is related to non-homologous end-joining and genome stability in *Aspergillus niger*. *Curr Microbiol* **62**, 1342–1346.
- Zhang, J., Chiodini, R., Badr, A. & Zhang, G. (2011b). The impact of next-generation sequencing on genomics. *J Genet Genomics* **38**, 95–109.
- Zhang, Y., Zhao, J., Fang, W., Zhang, J., Luo, Z., Zhang, M., Fan, Y. & Pei, Y. (2009). Mitogen-Activated Protein Kinase hog1 in the Entomopathogenic Fungus *Beauveria bassiana*

References

- Regulates Environmental Stress Responses and Virulence to Insects. *Appl Environ Microbiol* **75**, 3787–3795.
- Zhao, Z., Liu, H., Wang, C. & Xu, J.-R. (2013).** Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. *BMC Genomics* **14**, 274.
- Zheng, A., Lin, R., Zhang, D., Qin, P., Xu, L., Ai, P., Ding, L., Wang, Y., Chen, Y. & other authors. (2013).** The evolution and pathogenic mechanisms of the rice sheath blight pathogen. *Nat Commun* **4**, 1424.
- Zhou, P., Zhang, G., Chen, S., Jiang, Z., Tang, Y., Henrissat, B., Yan, Q., Yang, S., Chen, C.-F. & other authors. (2014).** Genome sequence and transcriptome analyses of the thermophilic zygomycete fungus *Rhizomucor miehei*. *BMC Genomics* **15**, 294.
- Zhou, Y.-H., Zhang, Y.-J., Luo, Z.-B., Fan, Y.-H., Tang, G.-R., Liu, L.-J. & Pei, Y. (2012).** Selection of optimal reference genes for expression analysis in the entomopathogenic fungus *Beauveria bassiana* during development, under changing nutrient conditions, and after exposure to abiotic stresses. *Appl Microbiol Biotechnol* **93**, 679–685.
- Znameroski, E. A., Coradetti, S. T., Roche, C. M., Tsai, J. C., Iavarone, A. T., Cate, J. H. D. & Glass, N. L. (2012).** Induction of lignocellulose-degrading enzymes in *Neurospora crassa* by cellodextrins. *Proc Natl Acad Sci*.

**Du génome au transcriptome pour la caractérisation des réseaux de régulation
contrôlant l'expression d'enzymes hydrolytiques chez un champignon d'intérêt
industriel**

Agustina LLANOS

Directeur : Jean-Luc PARROU

Co-directeur : Virginie NEUGNOT-ROUX

Introduction

Les enzymes hydrolytiques sécrétées par les champignons filamenteux ont trouvé des applications dans différentes industries, notamment dans l'alimentation humaine et animale, la fabrication du papier, ainsi que l'industrie textile. *Talaromyces versatilis* (précédemment nommé *Penicillium funiculosum*) peut, quant à lui, sécréter une grande variété d'enzymes capables de dégrader les polysaccharides complexes de la paroi des cellules végétales. Un cocktail d'enzymes produit par fermentation liquide de la souche *T. versatilis* IMI 378536 est actuellement commercialisé par Adisseo sous le nom de Rovabio™. Il est utilisé comme additif dans l'alimentation animale afin d'augmenter la valeur nutritionnelle des céréales chez les porcs et les volailles, et ainsi améliorer la performance des animaux. Adisseo s'est engagé depuis plusieurs années dans le développement et la recherche sur ce cocktail et le champignon qui le produit, pour améliorer sa performance.

Talaromyces versatilis a un important potentiel pour la production d'enzymes, qui, en plus de l'alimentation animale, pourraient avoir un intérêt pour d'autres applications. Il est en particulier impératif d'avoir une meilleure compréhension des mécanismes qui contrôlent l'expression des gènes codant les protéines composant ce cocktail et plus spécifiquement les glycosides hydrolases. C'est dans ce cadre que s'inscrit cette thèse, dont les principaux objectifs sont :

- Analyser la séquence génomique de *T. versatilis* afin de cataloguer les gènes codant les glycosides hydrolases, les protéines impliquées dans la voie de sécrétion et les facteurs de transcription potentiellement impliqués dans le contrôle de l'expression des gènes des deux premiers groupes

- Optimiser les outils moléculaires afin de faciliter la manipulation génétique de *T. versatilis* pour construire des mutants de délétion qui serviront à l'étude fonctionnelle des gènes.
- Analyser le transcriptome de *T. versatilis* ainsi que de mutants ayant des contextes génétiques différents, exposés à des différentes sources de carbone afin de caractériser les réseaux de régulation impliqués dans l'expression des gènes codant des enzymes hydrolytiques

Résultats

1. Croissance de *T. versatilis* en présence de différentes sources de carbone en comparaison avec d'autres champignons d'intérêt industriel

La croissance de *T. versatilis* a été évaluée en milieu solide et liquide. La croissance en milieu solide a permis de comparer le comportement de *T. versatilis* face à différentes sources de carbone. Du milieu gélosé minimum (TvMM) supplémenté avec différentes sources de carbone, a été inoculé au centre avec cent conidies de *T. versatilis*. Un suivi de croissance à 30°C a été effectué quotidiennement. *T. versatilis* a montré une capacité remarquable à pousser en présence de sources de carbone complexes, comme la cellulose (Avicel) ou la paille de blé, même si la croissance s'avère lente. Ceci suggère que *T. versatilis* peut produire toutes les enzymes nécessaires pour la dégradation de ce type de substrat pour ensuite incorporer les saccharides plus simples pour la production de biomasse et d'énergie.

La croissance en milieu liquide a permis de tester 2 sources d'azote, l'ammonium et le nitrate. Le temps de génération de *T. versatilis* s'est avéré élevé avec les deux sources d'azote mais tout de même plus court avec l'ammonium. Si les champignons sont capables d'utiliser une source organique d'azote, tous ne peuvent pas utiliser une source inorganique. Pour que le nitrate soit incorporé, la cellule doit d'abord le convertir en nitrite puis en ammonium. L'ammonium peut être combiné avec des acides organiques, généralement l'acide glutamique ou aspartique, pour donner des aminoacides. L'ammonium est, donc, une source d'azote plus convenable.

2. Annotation du génome de *T. versatilis*

Le génome de *T. versatilis* a été séquencé par Adisseo et une annotation automatique a été effectuée. 18 670 gènes ont été identifiés par annotation automatique et sont répartis sur 751 contigs. Cependant, cette annotation contient un nombre considérable d'erreurs, comme le manque de certains ORFs, une mauvaise délimitation d'ORFs, des exons mal placés ou manquants, etc. Afin de pouvoir développer des stratégies d'étude de la génomique fonctionnelle il est nécessaire de corriger ces erreurs par une annotation manuelle et experte du génome de *T. versatilis*.

Lors de l'analyse et la comparaison des séquences avec celles d'autres champignons filamenteux, la similarité avec *Talaromyces marneffe* et *Talaromyces stipitatus* a été mise en évidence. En plus de l'importante similarité de séquences avec ces deux champignons, des régions de synténie ont aussi été détectées. Dans ces régions, la position des gènes est très bien conservée, ce qui indique la faible ou nulle quantité d'événements de réarrangement,

mais aussi la proximité phylogénétique de ces espèces. Ceci supporte la décision de ré-classifier de notre champignon d'intérêt, *T. versatilis*, du genre *Penicillium* au genre *Talaromyces*.

L'annotation manuelle du génome de *T. versatilis* a été focalisée sur 3 groupes de gènes : les glycoside hydrolases (GH), les facteurs de transcription impliqués dans l'expression des glycoside hydrolases et les gènes impliqués dans le processus de sécrétion.

Les enzymes hydrolytiques font partie d'un groupe plus large d'enzymes : les enzymes actives sur les polysaccharides (CAZymes – www.cazy.org), responsables de l'assemblage, la modification et la dégradation des oligo- et polysaccharides. Ces enzymes sont classées en cinq groupes différents : les glycoside hydrolases (GH), les glycoside transférases (GT), les polysaccharide lyases (PL), les polysaccharides estérases (CE) et les activités auxiliaires (AA). Les GHs, PLs, CE, et AAs sont impliquées dans la dégradation des différentes polymères qui font partie de la paroi végétale.

Afin de saisir le potentiel génétique de *T. versatilis* pour la dégradation des polysaccharides de la paroi végétale, 32 familles de GHs ont été analysées dans son génome. C'est ainsi que 170 gènes ont été inventoriés, codant pour une grande variété d'enzymes, dont les activités sont très diverses mais complémentaires pour la dégradation des substrats lignocellulosiques. Certaines familles sont composées d'un nombre important des gènes, comme la famille GH3 qui contient 24 gènes, alors que d'autres sont composées d'un seul gène, comme la famille GH6. Un nombre considérable de gènes codent des hemicellulases, notamment des arabinofuranosidases des familles GH54 et GH62 et des xylanases de la famille GH11. Cependant, l'expression de ces gènes peut être soumise à des contrôles différents, et ils peuvent donc être exprimés dans des conditions distinctes.

Des gènes codant huit protéines hydrophobiques de binding et deux swollenines ont aussi été répertoriés dans le génome de *T. versatilis*. La fonction de ces protéines n'est pas entièrement comprise à ce jour, mais elles pourraient être impliquées de manière indirecte dans la dégradation des polysaccharides en augmentant la capacité d'hydrolyse des autres enzymes. La présence de gènes codant pour ces protéines auxiliaires, en plus des GHs, met encore une fois en évidence le grand potentiel de ce champignon pour une décomposition efficace des divers polymères des plantes.

Le deuxième groupe de gènes analysés correspond aux facteurs de transcription, et particulièrement ceux impliqués dans le contrôle de l'expression des gènes codant des enzymes hydrolytiques. Certains de ces facteurs ont été identifiés et caractérisés chez un certain nombre de champignons filamenteux. XlnR, AraR, ClrA, ManR, PacC, ACEII et AreA ont été décrits comme étant des régulateurs positifs, alors que CreA et ACEI sont les répresseurs les plus importants connus jusqu'à présent. Tous ces facteurs de transcription appartiennent à la famille des protéines de fixation à l'ADN avec de l'affinité pour le zinc. Des gènes codant pour des homologues de la plupart de ces facteurs ont été trouvés dans le génome de *T. versatilis*. Dans la plupart des cas, les séquences de ces facteurs de transcription ne présentent pas des divergences importantes avec ceux connus dans d'autres champignons filamenteux. Les homologues de *creA*, *creB* et *creC*, impliqués dans la répression catabolique, ont été identifiés. Les gènes codant pour les activateurs qui répondent à la présence de xylose, arabinose, cellulose et mannobiose ont aussi été identifiés : *xlnR*, *araR*, *clrA* et *manR*, respectivement. De plus, les huit gènes impliqués dans le « Pal pathway », qui est régulé par le pH de l'environnement, ont été retrouvés : *palA*, *palBory*, *palC*, *palF*, *palH*, *pall*, *pacC* et *vps32/snf7*. Un homologue du gène *areA*, qui code

pour un facteur de transcription qui contrôle l'expression des cellulases en réponse à la disponibilité d'azote, a été identifié. Enfin, un homologue du gène *hacA*, responsable de la réponse à la présence des protéines mal repliées (*Unfolded Proteins Response*), est aussi présent dans le génome. Huit de ces gènes ont eu besoin d'être manuellement corrigés.

Le troisième volet de l'analyse génomique a porté sur l'étude de la voie de sécrétion dont l'efficacité est essentielle pour des tels organismes du à que le potentiel de croissance repose sur la libération d'enzymes pour la dégradation des substrats. Dans des conditions d'induction, une grande quantité de protéines est transloquée à l'intérieur du réticulum endoplasmique ce qui nécessite une augmentation de l'efficacité de repliement et de transport de protéines. Un nombre important de gènes est impliqué dans ce processus qui comprend l'UPR et l'ERAD (*ER-associated protein degradation*). Les gènes connus chez *A. nidulans* ont été comparés à ceux de *T. versatilis*, et le niveau de similarité trouvé est remarquable, même s'il s'agit des champignons qui ne sont pas très proches dans la phylogénie. Ces résultats nous mènent à penser que la voie de sécrétion est extrêmement bien conservée, malgré les événements de spéciation qui ont du avoir lieu entre ces deux organismes.

3. Etude du transcriptome de *T. versatilis*

3.1. Séquençage du transcriptome de *T. versatilis*

Dans un premier temps, une étude exploratoire du transcriptome, en utilisant la technique de séquençage d'ARN, a été faite. Pour cela, le mycélium a été exposé a trois conditions différentes : glucose (condition de référence – Glucose 48h), paille de blé

(condition d'induction – Straw 24h), et paille de blé et glucose simultanément (Straw + Glucose 5h). Cette étude a permis d'identifier les gènes différentiellement exprimés entre les différentes conditions.

Un important changement dans le transcriptome de *T. versatilis* a été observé entre les deux conditions les plus extrêmes : Glucose 48h et Straw 24h. Dans cette expérience, le mycélium est transféré d'un milieu avec glucose, qui est un substrat facile de métaboliser et répresseur de la plupart des systèmes d'utilisation de substrats alternatifs, à un milieu avec de la paille de blé essentiellement constituée de lignocellulose, qui est un substrat beaucoup plus complexe et qui nécessite d'être dégradé afin que des molécules plus petites puissent être incorporées par le champignon et métabolisées.

C'est ainsi que 48 % de gènes codant pour des enzymes hydrolytiques ont été induits lors de l'exposition à la lignocellulose, beaucoup d'entre eux codent pour des enzymes qui possèdent une activité similaire. Dans certains cas, un nombre important des gènes appartenant à une même famille CAZy ont été induits. Le fait que des enzymes avec une activité similaire soit induites peut être justifié par l'effet synergique entre elles. Le résultat le plus remarquable concerne toutefois la masse totale d'ARN correspondant aux gènes codant des GHs. En effet, lors que les 170 gènes ne donnaient que 0.1 % du transcriptome sur glucose, l'induction de ces gènes sur la paille de blé entraîne une accumulation massive de leurs transcrits, qui représente alors 13 % du total du transcriptome. Ceci suggère que le champignon fait un grand investissement pour faire face à la nouvelle condition.

Un deuxième grand changement est observé lors de l'addition de glucose dans la troisième partie de cette expérience. Les gènes codant pour des cellulases et des

hemicellulases sont réprimés, même si la répression n'est pas aussi forte que lorsque le mycélium était en croissance sur glucose.

Cette étude a aussi permis d'identifier d'autres gènes induits par la présence de la lignocellulose dans le milieu, comme les swollenines et les protéines hydrophobiques de binding, ainsi qu'un grand nombre des protéines de transport. La fonction des protéines hydrophobiques et des swollenines n'est pas encore bien comprise, mais la forte induction de ces gènes peut être liée à leur participation indirecte dans la dégradation des substrats complexes. Des études supplémentaires seront nécessaires pour décrypter leur fonction et implication dans ce processus.

Parmi les facteurs de transcription connus impliqués dans le contrôle des glycosides hydrolases, seuls *xlnR* et *clrA* ont été différenciellement exprimés. D'autre part, *hacA*, impliqué dans la régulation de la réponse UPR, est aussi induit mais plus modestement, avec un FC de 3.8. En plus de ces gènes, 11 autres gènes codant des facteurs de transcription putatifs ont été identifiés, et présentent une forte induction quand le champignon est transféré dans le milieu supplémenté avec la paille de blé. Tous codent pour des protéines qui contiennent un domaine de binding à l'ADN. Leur véritable fonction n'est pas connue, mais ils pourraient être impliqués dans le control de l'expression des gènes codant pour des GHs, et contribuer ainsi à leur induction.

Dans le cadre direct des travaux de doctorat, nous avons pu également profiter de la nouvelle annotation, fournie par la plateforme « *Deep-seq* » de l'Université de Nottingham, pour faire un analyse de « *GO term enrichment* », au-delà des simples familles de GHs ou TFs. Les GO (*Gene Ontology*) terms activés lors de l'exposition à la lignocellulose incluent le catabolisme de polysaccharides, l'activité cellulase, l'activité hydrolytique et xylanase, le

métabolisme de la pectine et le transport. Toutes ces fonctions sont nécessaires pour la dégradation des polymères présents dans la paroi végétale. La présence de lignocellulose a aussi réprimé des GO terms comme ceux de synthèse de ribosomes, maturation des ARN ribosomiaux, synthèse des aminoacides, fonctions mitochondriales, synthèse d'ATP, transport de protons et développement du mycélium. Par conséquent, le champignon semble être dans un état « stable » mais sans croissance et sans développement. Cette situation inattendue est complètement inversée dès l'ajout de glucose dans le milieu. La présence d'un substrat aussi complexe comme la paille de blé pourrait générer un stress auquel la cellule répond par la répression de fonctions fondamentales.

3.2. Sélection des gènes de référence pour la normalisation de données de RT-qPCR

Une étape critique pour l'étude de l'expression génique par RT-qPCR est la normalisation des données, l'une des stratégies la plus utilisée s'appuie sur la disponibilité de gènes de référence fiables. Nous avons identifiés et validés une sélection de gènes de référence pour la quantification relative de l'expression des gènes chez *T. versatilis*.

Une première sélection de gènes de référence potentiels a été réalisée *in silico* à partir de données de transcriptomique. Une douzaine de candidats a été analysé par RT-qPCR sur plus de 30 conditions de culture. En utilisant geNorm, nous avons montré que la plupart de ces gènes avaient des niveaux de transcrits stables dans la plupart des conditions, dont des conditions de stress, une grande variété de sources de carbone et pendant la germination des conidies. Nous avons montré aussi que toute combinaison de trois de ces gènes de référence conduit à un biais de normalisation minimale quelque soit la condition.

Pour aller au-delà de *T. versatilis*, nous avons analysé la stabilité de ces gènes de référence ainsi que des gènes de référence classiquement utilisés, tels que la β -tubuline ou l'actine, dans 18 champignons filamenteux phylogénétiquement éloignés, exposés à des conditions expérimentales différentes. Bien que cette analyse a démontré que chacun des gènes choisis montre des cas d'induction ou de répression sporadiques, une classification hiérarchique de tous ces gènes a permis l'identification d'un groupe de 6 gènes. Ce groupe présente de faibles changements d'expression et aucun biais de régulation sur l'ensemble des conditions. Il comprend les gènes *ubcB*, *SAC7* et *Fis1*, ainsi que *sarA*, *TFC1* et *UBC6*. Par conséquent, nous proposons un ensemble de six gènes qui peuvent être utilisés comme gènes de référence pour la normalisation des données de RT- qPCR dans n'importe quel domaine de la biologie des champignons. Cependant, la stabilité de l'expression de ces gènes de référence doit être testé et validé de façon systématique avant leur utilisation. L'utilisation de la moyenne géométrique d'au moins trois des meilleurs gènes est fortement conseillé, ce qui permettrait de minimiser le biais dans la normalisation et l'obtention de conclusions biologiques fiables.

3.3. Analyse de l'expression des GH par RT-qPCR

Afin de compléter l'étude du transcriptome, la technique de PCR quantitative a été utilisée pour faire une analyse en profondeur de l'expression de certains gènes d'intérêt, incluant des gènes codant pour des cellulases, hemicellulases et des protéines auxiliaires. L'objectif de ce travail était d'analyser l'expression de ces gènes dans un éventail plus large des conditions. Le mycélium de *T. versatilis* a été exposé à des substrats complexes, comme la paille de blé ou le xylan, mais aussi à des monosaccharides et disaccharides. Les sucres

choisis pour cette expérience sont ceux produits lors de la dégradation de polymères comme la cellulose ou l'arabinoxylan. De cette façon nous pouvons décortiquer la réponse du champignon à ces différents stimuli. Nous avons aussi inclus des conditions de stress, comme des pH extrêmes ou le manque d'azote ou de carbone.

L'expression de ces gènes d'intérêt dans toutes les conditions a d'abord été analysée avec une classification hiérarchique ascendante (HAC). Cela a permis une première classification des gènes par rapport à leur comportement général. Cependant, cette analyse ne nous a pas permis d'identifier les cas de co-régulation en réponse à un substrat donnée, c'est-à-dire lorsque des gènes sont induits ou réprimés de façon similaire en réponse à un même substrat. Nous avons donc étudié la corrélation entre les gènes et ainsi regroupés les gènes ayant un comportement similaire dans toutes les conditions. Les groupes ont été déterminés en fonction de la valeur du coefficient de corrélation. Cinq groupes ou classes ont pu ainsi être identifiés. De façon très intéressante, chacune des classes est majoritairement composée de gènes codant pour un même type d'enzymes :

- Classe 1 : les cellulases
- Classe 2 : les hemicellulases qui ont des très hauts niveaux d'expression
- Classe 3 : les hemicellulases avec des niveaux d'expression bas
- Classe 4 : les β -xylosidases
- Classe 5 : la protéine hydrophobique

La similarité dans l'expression des gènes appartenant à la même classe peut indiquer qu'ils partagent des systèmes de régulation grâce à des éléments présents dans leur région promotrice. Une comparaison de ces régions pourrait être envisagée pour identifier les séquences conservées entre les différents promoteurs.

4. Effet de la délétion des facteurs de transcription clés dans l'expression des gènes codant pour des GHs

L'analyse génétique des champignons filamenteux, et en particulier de *T. versatilis*, peut être mise en péril par le manque d'outils permettant l'obtention de souches mutantes. Le nombre limité de marqueurs de sélection disponibles et la faible efficacité de la recombinaison homologue par rapport à l'intégration ectopique de fragments linéaires d'ADN est une autre inquiétude lors de la préparation des souches modifiées génétiquement. Aussi, une méthode permettant la délétion de gènes, développée chez *A. niger*, a été adaptée à *T. versatilis*. Ce système permet de s'affranchir de l'utilisation d'une souche déficiente NHEJ qui est généralement génétiquement instable. Potentiellement, il permet aussi de recycler le marqueur de sélection et ainsi déléter successivement plusieurs gènes dans la même souche.

L'utilisation de ce système chez *T. versatilis* a d'abord nécessité l'obtention d'une souche délétée dans le locus *pyrG*. Ensuite, le gène *pyrG* d'*A. oryzae* a été testé chez *T. versatilis* afin de vérifier la possibilité de l'utiliser comme marqueur de sélection. Par la suite, le plasmide pC3, qui contient une copie du gène *pyrG* de *A. oryzae* et un site de multi-clonage, a été utilisé pour déléter deux facteurs de transcription chez *T. versatilis*. Cette méthode de délétion devrait être efficace avec d'autres espèces et genres de champignons filamenteux, à condition d'avoir une souche de délétion *pyrG* et une complémentation fonctionnelle par le gène *pyrG* présent dans le plasmide.

Les deux mutants de délétion obtenus - $\Delta xlnR$ et $\Delta clrA$ – nous ont permis d'initier l'étude de la régulation des gènes codant pour des GHs chez *T. versatilis*. Les résultats de qPCR ont en effet montré que l'expression des gènes dans les classes 2 (les principales hemicellulases)

et 4 (les β - xylosidases) est sérieusement affectée par l'absence de XlnR, même si ces gènes sont encore induits en absence du régulateur xylanolytique. Nous avons également observé que l'expression de ces gènes n'est pas affectée lorsque la souche mutante est exposé à de l'arabinose. Lorsque nous avons analysé l'effet de l'absence de XlnR dans les niveaux d'expression des gènes de la classe 1 (les cellulases) nous avons trouvé que l'expression n'est pas réduite en réponse aux polysaccharides. Au contraire, certains d'entre eux ont été encore plus induits dans la souche mutante que dans la souche sauvage. XlnR est également impliqué dans le contrôle de l'expression des gènes de la classe 3. Ces gènes ont été plus induits dans la souche mutante que dans la souche sauvage, mais seulement en présence du xylan. Pour les autres substrats testés, les niveaux de transcrits ont été plus faibles chez la souche mutante. Le gène qui code pour l'hydrophobine montre un profil d'expression très particulier qui est éloignée des autres gènes étudiés. La forte induction observée dans la souche sauvage en présence de lignocellulose est fortement réduite dans le mutant de délétion $\Delta xlnR$, ce qui suggère que XlnR est en effet impliqué dans l'activation de la transcription de ce gène. Les expériences préliminaires effectuées avec la souche $\Delta xlnR$ ont montré que XlnR est nécessaire pour l'induction optimale des gènes codant pour des hemicellulases : ceci est cohérent avec les résultats qualitatifs obtenus avec le test du rouge Congo, qui a montré que la même souche a une capacité réduite pour la dégradation du xylan par rapport à la souche sauvage.

En ce qui concerne la souche *T. versatilis* $\Delta clrA$, un test de dégradation de polysaccharides a été réalisé. L'expression des GHs a été induite en exposant la souche mutante à de la paille de blé, qui est typiquement un inducteur très puissant, mais aucune activité de cellulase a été observé dans le test du rouge Congo. Plus intéressant encore,

aucune activité xylanase n'a été observée non plus. Ceci indique que le régulateur ClrA joue un rôle important dans l'activation des gènes codant pour la cellulase et xylanase. Pour confirmer ces résultats, des expériences supplémentaires sont nécessaires, notamment une étude des niveaux de transcrits des différentes classes des gènes d'intérêt dans ce contexte génétique.

Conclusions générales

L'objectif principal de cette thèse était de caractériser les réseaux de régulation qui contrôlent l'expression des gènes codant pour les enzymes impliquées dans la dégradation des polysaccharides végétaux. Nous avons analysé la croissance de *T. versatilis* avec différentes sources de carbone. Ensuite, nous avons utilisé une approche génomique pour identifier les gènes impliqués dans la production et la sécrétion d'enzymes hydrolytiques, et une approche transcriptomique pour étudier le profil d'expression des gènes codant pour des enzymes hydrolytiques. Ainsi, les faits marquants sont :

Sur les aspects physiologiques :

- *T. versatilis* montre un bon taux de croissance sur les sources de carbone complexes et préfère ammonium comme source d'azote
- L'analyse du génome de *T. versatilis* a confirmé que ce champignon a le potentiel génétique pour faire face à un style de vie saprophyte en présence de sources de carbone complexes qui doivent être dégradées en monomères avant d'être incorporées et métabolisées

Sur le transcriptome :

- L'étude du transcriptome par RNA-seq de *T. versatilis* a révélé l'énorme réorganisation du transcriptome lorsque le champignon, préalablement exposé au glucose, est ensuite transféré dans un milieu contenant de la lignocellulose
- Différents profils d'expression de gènes codant pour des cellulases et hemicellulases ont été révélés en réponse à divers inducteurs
- La souche *T. versatilis* $\Delta xlnR$ a montré une expression réduite des gènes codant pour des hemicellulases, ce qui suggère que XlnR est impliqué dans le contrôle de l'expression de ces gènes

Sur les méthodologies mises au point :

- La sélection de nouveaux gènes de référence putatifs pour la normalisation des données de RT- qPCR chez les champignons filamenteux
- La mise au point d'une méthode de transformation avec un gène marqueur recyclable a été adaptée à *T. versatilis* et utilisée pour la manipulation génétique de cette souche

