7-1-2010

Genome architecture in the fungal kingdom

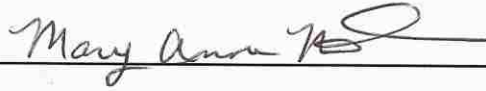Antonio Diego Martinez

Diego Antonio Martinez
Candidate

Biology
Department

This dissertation is approved, and it is acceptable in quality
and form for publication on microfilm:

Approved by the Dissertation Committee:

_____ , Chairperson

_____

_____

_____

_____

_____

_____

_____

GENOME ARCHITECTURE IN THE FUNGAL KINGDOM

BY

DIEGO ANTONIO MARTINEZ

B.S., Biology, The University of New Mexico, 2001
B.S., Chemistry, The University of New Mexico, 2001

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

**Doctor of Philosophy
Biology**

The University of New Mexico
Albuquerque, New Mexico

July, 2010

DEDICATION

I dedicate this work to my daughter Enorah, who motivated me to focus on my work so I can finish and pay more attention to her. This is also dedicated to my wife Sarah, who saw the diamond in the rough that I was and polished me to a brilliant sparkle. Through her belief in me and with her support I am able to achieve all my goals.

ACKNOWLEDGEMENTS

My career has had many twists and is not a normal path. This means that there were many along the way who helped me get to where I am today.

First, I would like to thank my academic parents who put up with me and supported me. Most of all I would like to thank Mary Anne Nelson who gave me my start in the Neurospora Genome Project and introduced me to the world of bioinformatics and also as my graduate advisor and mentor. I can never repay her for the support and mentorship she has given me. I will continue my career with her as my model of what a mentor should be. Without Maggie I would not know how to follow my heart, nor would I have been paid! To Don I would not know the command line or what *Phanerochaete chrysosporium* actually looks like!

In my career I have the wonderful opportunity to work with many collaborators who have contributed significantly to my development in science. First I would like to thank Randy Berka and Dan Cullen. These two people helped me with my first, second and third papers in which I was the first author. From them I have learned the ropes of publishing science. I would also like to thank Mikko Arvas and Markku Saloheimo, the geniuses of *Trichoderma reesei*. Their assistance on the *T. reesei* genome project was essential! I also thank the JGI for the boot camp experience in genomics and all the people there for their support and guidance and mentorship in genome sciences. At LANL I would like to thank Gary Xie and Jean Challacombe for their support and mentorship as well, and all their help at the Jamboree and teaching me annotation.

I thank the undergraduates who have helped me on this project. With out them I would still be figuring out how to make pathologic work and be checking jobs on pequena. Thanks to Christine Chee, Charles Sanchez and Joe Kunkel for helping me start and finish all the work in chapter 3 of this dissertation! Without Daniel Chee I would have gone to coffee alone for 10 months! Thank you all!

I thank Sushmita Roy and Osorio Meirelles who mentored me in Computer Science and Statistics. I would know nothing about modeling without them and I hope to collaborate with them soon.

GENOME ARCHITECTURE IN THE FUNGAL KINGDOM

BY

DIEGO ANTONIO MARTINEZ

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

**Doctor of Philosophy
Biology**

The University of New Mexico
Albuquerque, New Mexico

July, 2010

GENOME ARCHITECTURE IN THE FUNGAL KINGDOM

By

Diego Antonio Martinez

B.S., Biology, The University of New Mexico, 2001
B.S., Chemistry, The University of New Mexico, 2001
Ph. D., Biology, The University of New Mexico, 2010

ABSTRACT

Previous studies have suggested that the location of genes in genomes is not random; instead they may be organized in a way that is beneficial to cellular processes and the organism. While a few studies have investigated the organization of genes on a whole genome scale, they were limited in the functions of genes used in the search and in the number and type of genomes searched. With the recent explosion of available fungal genomes and tools to automatically annotate many genes in a short period of time, it is now possible to obtain a global view of the level of clustering in the genomes of an entire kingdom. To find gene clusters in many genomes, we have constructed a robust and flexible algorithm that runs in trivial time. In parallel, we have annotated 72 fungal genomes using four automated annotation tools that provide information about protein function, protein targeting, involvement in biochemical pathways and paralogous gene families. We used the clustering algorithm to search for clusters from the four annotation categories. We discovered that all the genomes contained clusters of related genes, and that in several cases the clusters included genes involved in processes that were specific to the species in which they are found. This has dramatically expanded our knowledge of both the types of clusters and the number of genomes known to contain clusters. This study has generated

information that will assist researchers in addressing many questions central to molecular and cell biology as well as evolutionary studies. To this end, the thousands of clusters we have discovered are available for download at kiddomics.com/.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 3

INTRODUCTION:

ADAPTIVE GENE TRANSLOCATIONS IN THE FUNGAL KINGDOM

Fungi in Genomics and Science

Fungi have long enjoyed a central role in biological research. From the one-gene-one-enzyme hypothesis (developed in *Neurospora crassa*) to confirming whole genome duplications[1], fungi have been important experimental organisms, key to establishing basic cause and effect relationships in molecular biology. This trend continues today, and is particularly important in the current genomic era.

Some consider the era of whole genome studies as beginning with fungi, and indeed the first sequenced eukaryote was a fungus. The genome of *Saccharomyces cerevisiae* was published in 1995, and this important yeast model system continues to be a workhorse in genomic science[2]. In fact, more fungal genomes have been sequenced or are being sequenced than for any other kingdom except prokaryotes [me and you 2010]. There are two reasons for this emphasis on analysis of fungal genomes. First, fungi have genomes of approximately 40 megabases, and they average roughly 10,000 protein coding genes. Second, they tend to have few repetitive elements, which makes whole genome assembly easier[3-5]. This is in contrast to the human genome, which is more than three gigabases in size and which contains an estimated 25,000 genes[6-8]. In addition, the speed at which fungi grow and reproduce simplifies DNA preparation time and effort. These factors together have established fungi as the bacteria of eukaryotic molecular biology.

Several features of the new whole genomes began to stimulate interest in genome architecture (reviewed in Hurst et al. 2004 [9]). At the turn of the century, several

experiments [9-12] indicated that genomes are not organized in a random fashion. While some of that analysis has since been questioned, additional research in other organisms supported similar non-random organization. The mounting examples pointed to potentially widespread evolution by adaptive gene relocation.

How Do Clusters Form?

In bacteria there is a mechanism to directly support clusters of related genes, i.e., operons with genes that must be transcribed together (genes are transcribed on a single polycistronic mRNA, which is translated to produce multiple proteins). Examples include the well known *lac* (*lactose*) operon [10] Another interesting example that parallels findings in our studies involves the cellulosome genes (encoding the proteins that form a pseudoorganelle of cellulase-degrading proteins in bacteria), which are often found in one or more operons[11]. In contrast, eukaryotes do not generally have operons (one exception being the model organism *Caenorhabditis elegans*). This raises the question as to what keeps gene clusters together in eukaryotes. Some speculate that the gene clusters were transferred from foreign organisms, and that the genomic features of the region are necessary for the survival of the genes, which keeps them from drifting [12].

While maintenance of gene clusters remains an open question, formation is a bit more tractable. The evidence comes from the unexpected instability of the genome and the amount of rearrangements seen. Analysis of changes on a whole genome scale became common as new algorithms were created to detect *synteny*, the level of chromosomal conservation in gene content between two species. Comparative analysis between species with respect to synteny also supported the finding that structural alterations are the norm in evolution [13] Additionally, small reordering events in gene order may be fairly common [14,

15]. These studies established that genomes are dynamic; the possibility that genome architecture can adapt to a more advantageous configuration no longer seemed improbable.

Pre-genomic Evidence

Were any examples of gene clusters found before whole genome sequences were available? Several investigations uncovered clusters of coexpressed genes involved in the same biochemical processes that were colocated in the genome. In filamentous fungi, the *qa* cluster (genes involved in quinolone metabolism) are adjacent to one another in the genome, a fact established more than 30 years ago [16] and confirmed in recent years with whole genome sequencing [17]. In yeast, the GAL locus contains three genes involved in galactose catabolism that are coregulated [18].

More clues in the pre-genomic era regarding the potential importance of genome organization came from the fungal kingdom. An important class of chemicals, both environmentally and economically, are produced by this kingdom. The genes that encode the protein machinery responsible for the production of secondary metabolites, important toxins, have for years been known to exist often in a tight cluster of genes it has been hypothesized that these genes were transfered into the genome as a unit from other organisms [19]. While the debate rages on, researchers on the side of horizontal gene transfer appear to be gaining traction in the argument as more examples are uncovered [20].

This anecdotal evidence in a variety of genomes suggests that some key genes are organized in a fashion that makes them more accessible to transcriptional machinery. How broad is this phenomenon? A first study in 2003 published by Lee and Sonnhammer

suggested that some of the genes encoding biochemical enzymes are clustered, including as many as 98% of the biochemical pathways in the yeast *Saccharomyces cerevisiae* [21]. A later study analyzing clusters of genes with related Gene Ontology terms found similar results [22]. Still, these studies were performed using only the classic model organisms (MOD), and were limited in the percentage of the genome that was analyzed. The statistical model used in the former study is not appropriate for wide application due to the dependency on chromosomal length; also, the former study uses annotation types not appropriate outside the MODs (due to the sparse nature of the terms). Questions still remained as to how common the phenomenon of clustering of genes is and how rapidly the clusters may be changing across species.

Broad Analysis and Clues to Formation

In 2008 we addressed several of these questions in the analysis of the genome sequence of *Trichoderma reesei* [3](Chapter 1 of this dissertation). By analyzing a non-MOD organism, we added another organism to the growing list of species that contain non-random genome organization. More importantly, we found clues as to the potential causes of the clustering phenomenon. We discovered that a sizable portion of the genes producing carbohydrate active enzymes (CAZy) [23] are clustered in the genome. Additionally, we found that the mixture of genes that are clustered contributes to the ability of *T. reesei* to degrade dead plant material. This gives clues about what causes gene clustering, as it ties a large gene family that is key for a particular *T. reesei* adaptation to a potentially favored architecture.

Chapter 1 also breaks ground in discovering links to reorganization and gene functions. This was one of the first analyses to compare synteny and gene clustering. We found that the CAZy clusters were predominantly in NSBs (non-syntenic blocks) as compared to other

4

closely related fungi. Adding to this finding, the fact that few of the CAZy genes are from the same family within a cluster (indicating reduced sequence similarity), implicates gene movements in the formation of the clusters over gene duplications. This expands on a previous report that a small gene cluster in *S. cerevisiae* involved in allantoin metabolism was formed by adaptive gene relocation [24]. Our finding is key to understanding formation of clusters.

Still, how quickly do these changes take place? In our 2009 study [25] (Chapter 2:Addendum of this dissertation), we had the opportunity to dig through the genome of another important biomass degrading fungus, the basidiomycete *Postia placenta*. This is one of the first genome assemblies that preserved the separate haplotypes that make up the diploid genome. This provided us the fortunate chance to investigate the connection between genome rearrangements and gene function on a short time scale.

To discover areas that might contain rearrangements, we constructed a custom algorithm to detect differences in the order of genes or sections of genome that might be caused by genome rearrangements. We excluded regions that contained assembly gaps on the boundaries of the candidate rearrangements, as those differences may simply reflect errors. Although we found more than 400 candidate regions between the two haplotypes that might contain genome rearrangements, we found no regions that could be included due to likely assembly errors. This suggests that in fungi, at the population level, genome rearrangements are not common. However, this is a severely limited study, making it difficult to draw major conclusions.

However, we did find interesting differences in the NSB areas containing clusters of potentially important biomass degrading genes, confirming the findings in Chapter 1. This is important, as the genes found in Chapter 2 (genome of *P. placenta*) have no similarity or relation to those of Chapter 1. This underscores the conclusion that in fungi, NSBs and gene clusters may be quite common.

As anecdotal evidence mounted, answering the question as to how broad the phenomenon of gene clustering is seemed a logical next step. With the abundance of fungal genomes sequenced, it was a rare opportunity to determine if this arrangement occurred across an entire kingdom. While the representation of sequenced genomes is somewhat weighted toward certain parts of the fungal tree of life, there are specimens from 4 of the major branches in the fungal kingdom.

In Chapter 3 we constructed an automated version of the manual steps in our 2008 study [3] (Chapter 1), which implemented a hypergeometric model of genome clustering. We then processed all fungal genomes available at the time, a total of 72. We noted that previous studies were limited in their search of clustering across different types of gene annotations. By limiting an investigation to one type of annotation, such as biochemical pathways, previous studies were unable to determine how layers of the genome were impacted. We attempted to combat this limitation by annotating most genomes in four different ways. First, by using Interpro [26] we were able to find clusters of genes that had the same function. Second, we annotated biochemical pathway enzymes and used the pathways as input to discover clusters; we were then able to expand on and compare our algorithm to previous work [21]. Third, we annotated genes by targeting location. Finally, we found paralogous clusters of gene products that spanned multiple genomes by applying the MCL algorithm

6

[27]. Sadly, we were unable (due to time constraints) to calculate MCL groups for all genomes and were only able to include 44 genomes in this clustering analysis.

We found that all genomes contained clusters of all four annotation categories, however with widely varying content. Interestingly, location targeting produced the largest clusters and most diverse groups of genes in the genome. In particular, genes that produced secreted proteins were the most intriguing clusters in fungal genomes, often encoding gene products that seemed to be necessary pieces of machinery tailored to the specific needs of the fungus for carbon acquisition, pathogenesis or other species-specific functions.

The paralogous gene groups also gave us the chance to test the movement of gene duplicates. We found that only 18%, on average, of gene duplicates remain clustered. This observation, in combination with the amount and type of clustering we found, suggests that gene movements are in fact common. This research lends weight to the hypothesis that certain gene movements are adaptive, and thus certain gene configurations are favored in evolution.

References

1. Kellis M, Birren BW, Lander ES: **Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae**. *Nature* 2004, **428**:617-624.

2. Goffeau A, Barrell B, Bussey H, Davis R, Dujon B, Feldmann H, Galibert F, Hoheisel J, Jacq C, Johnston M, Louis E, Mewes H, Murakami Y, Philippsen P, Tettelin H, Oliver S: **Life with 6000 genes**. *Science* 1996, **274**:546-&.

3. Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, Chapman J, Chertkov O, Coutinho PM, Cullen D, Danchin EGJ, Grigoriev IV, Harris P, Jackson M, Kubicek CP, Han CS, Ho I, Larrondo LF, de Leon AL, Magnuson JK, Merino S, Misra M, Nelson B, Putnam N, Robbertse B, Salamov AA, Schmoll M, Terry A, Thayer N, Westerholm-Parvinen A, Schoch CL, Yao J, Barbote R, Nelson MA, Detter C, Bruce D, Kuske CR, Xie G, Richardson P, Rokhsar DS, Lucas SM, Rubin EM, Dunn-Coleman N, Ward M, Brettin TS: **Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina)**. *Nat Biotech* 2008, **26**:553-560.

4. Cuomo C, Gueldener U, Xu J, Trail F, Turgeon B, Di Pietro A, Walton J, Ma L, Baker S, Rep M, Adam G, Antoniw J, Baldwin T, Calvo S, Chang Y, DeCaprio D, Gale L, Gnerre S, Goswami R, Hammond-Kosack K, Harris L, Hilburn K, Kennell J, Kroken S, Magnuson J, Mannhaupt G, Mauceli E, Mewes H, Mitterbauer R, Muehlbauer G, Munsterkotter M, Nelson D, O'Donnell K, Ouellet T, Qi W, Quesneville H, Roncero M, Seong K, Tetko I, Urban M, Waalwijk C, Ward T, Yao J, Birren B, Kistler H: **The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization**. *Science* 2007, **317**:1400-1402.

5. Galagan J, Calvo S, Borkovich K, Selker E, Read N, Jaffe D, FitzHugh W, Ma L, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen C, Butler J, Endrizzi M, Qui D, Ianakiev P, Pedersen D, Nelson M, Werner-Washburne M, Selitrennikoff C, Kinsey J, Braun E, Zelter A, Schulte U, Kothe G, Jedd G, Mewes W, Staben C, Marcotte E, Greenberg D, Roy A, Foley K, Naylor J, Stabge-Thomann N, Barrett R, Gnerre S, Kamal M, Kamvysselis M, Mauceli E, Bielke C, Rudd S, Frishman D, Krystofova S, Rasmussen C, Metzenberg R, Perkins D, Kroken S, Cogoni C, Macino G, Catcheside D, Li W, Pratt R, Osmani S, DeSouza C, Glass L, Orbach M, Berglund J, Voelker R, Yarden O, Plamann M, Seller S, Dunlap J, Radford A, Aramayo R, Natvig D, Alex L, Mannhaupt G, Ebbole D, Freitag M, Paulsen I, Sachs M, Lander E, Nusbaum C, Birren B: **The genome sequence of the filamentous fungus Neurospora crassa**. *Nature* 2003, **422**:859-868.

6. Stein LD: **Human genome: End of the beginning**. *Nature* 2004, **431**:915-916.

7. Lander ES, Linton LM, Birren B, et al.: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.

8. Venter JC, Adams MD, Myers EW, et al.: **The Sequence of the Human Genome**. *Science* 2001, **291**:1304-1351.

9. Hurst L, Pal C, Lercher M: **The evolutionary dynamics of eukaryotic gene order**. *Nat. Rev. Genet.* 2004, **5**:299-310.

10. Jacob F, Monod J: **Genetic regulatory mechanisms in the synthesis of proteins**. *J. Mol. Biol* 1961, **3**:318-356.

11. Doi R, Kosugi A: **Cellulosomes: Plant-cell-wall-degrading enzyme complexes**. *Nature Reviews Microbiology* 2004, **2**:541-551.

12. Khaldi N, Wolfe KH: **Elusive Origins of the Extra Genes in Aspergillus oryzae**. *PLoS ONE* 2008, **3**:e3036.

13. Eichler EE, Sankoff D: **Structural Dynamics of Eukaryotic Chromosome Evolution**. *Science* 2003, **301**:793-797.

14. Seoighe C, Federspiel N, Jones T, Hansen N, Bivolarovic V, Surzycki R, Tamse R, Komp C, Hulzar L, Davis R, Scherer S, Tait E, Shaw D, Harris D, Murphy L, Oliver K, Taylor K, Rajandream M, Barrell B, Wolfe K: **Prevalence of small inversions in yeast gene order evolution**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**:14433-14437.

15. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J, Benito-Gutierrez E, Dubchak I, Garcia-Fernandez J, Gibson-Brown JJ, Grigoriev IV, Horton AC, de Jong PJ, Jurka J, Kapitonov VV, Kohara Y, Kuroki Y, Lindquist E, Lucas S, Osoegawa K, Pennacchio LA, Salamov AA, Satou Y, Sauka-Spengler T, Schmutz J, Shin-I T, Toyoda A, Bronner-Fraser M, Fujiyama A, Holland LZ, Holland PWH, Satoh N, Rokhsar DS: **The amphioxus genome and the evolution of the chordate karyotype**. *Nature* 2008, **453**:1064-1071.

16. Giles NH, Case ME, Baum J, Geever R, Huiet L, Patel V, Tyler B: **Gene organization and regulation in the qa (quinic acid) gene cluster of Neurospora crassa**. *Microbiol Rev* 1985, **49**:338-358.

17. Dean R, Talbot N, Ebbole D, Farman M, Mitchell T, Orbach M, Thon M, Kulkarni R, Xu J, Pan H, Read N, Lee Y, Carbone I, Brown D, Oh Y, Donofrio N, Jeong J, Soanes D, Djonovic S, Kolomiets E, Rehmeyer C, Li W, Harding M, Kim S, Lebrun M, Bohnert H, Coughlan S, Butler J, Calvo S, Ma L, Nicol R, Purcell S, Nusbaum C, Galagan J, Birren B: **The genome sequence of the rice blast fungus Magnaporthe grisea**. *Nature* 2005, **434**:980-986.

18. Bassel J, Mortimer R: **Genetic order of the galactose structural genes in Saccharomyces cerevisiae**. *J. Bacteriol* 1971, **108**:179-183.

19. Giles NH: **The Organization, Function, and Evolution of Gene Clusters in Eucaryotes**. *The American Naturalist* 1978, **112**:641-657.

20. Mallet L, Becq J, Deschavanne P: **Whole genome evaluation of horizontal transfers in the pathogenic fungus Aspergillus fumigatus**. *BMC Genomics* 2010, **11**:171.

21.   Lee JM, Sonnhammer EL: **Genomic Gene Clustering Analysis of Pathways in Eukaryotes**. *Genome Res.* 2003, **13**:875-882.

22.   Yi G, Sze S, Thon MR: **Identifying clusters of functionally related genes in genomes**. *Bioinformatics* 2007, **23**:1053-1060.

23.   Coutinho P, Henrissat B: **Carbohydrate-active enzymes: An integrated database approach**. In *Recent Advances in Carbohydrate Bioengineering*. Cambridge, United Kingdom: The Royal Scociety of Chemistry; 1999:3-14.

24.   Wong S, Wolfe KH: **Birth of a metabolic gene cluster in yeast by adaptive gene relocation**. *Nat Genet* 2005, **37**:777-782.

25.   Martinez D, Challacombe J, Morgenstern I, Hibbett D, Schmoll M, Kubicek CP, Ferreira P, Ruiz-Duenas FJ, Martinez AT, Kersten P, Hammel KE, Vanden Wymelenberg A, Gaskell J, Lindquist E, Sabat G, Splinter BonDurant S, Larrondo LF, Canessa P, Vicuna R, Yadav J, Doddapaneni H, Subramanian V, Pisabarro AG, Lavín JL, Oguiza JA, Master E, Henrissat B, Coutinho PM, Harris P, Magnuson JK, Baker SE, Bruno K, Kenealy W, Hoegger PJ, Kües U, Ramaiya P, Lucas S, Salamov A, Shapiro H, Tu H, Chee CL, Misra M, Xie G, Teter S, Yaver D, James T, Mokrejs M, Pospisek M, Grigoriev IV, Brettin T, Rokhsar D, Berka R, Cullen D: **Genome, transcriptome, and secretome analysis of wood decay fungus Postia placenta supports unique mechanisms of lignocellulose conversion**. *Proceedings of the National Academy of Sciences* 2009, **106**:1954-1959.

26.   Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJA, Zdobnov E: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites**. *Nucleic Acids Research* 2001, **29**.

27.   Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families**. *Nucl. Acids Res.* 2002, **30**:1575-1584.

CHAPTER ONE

GENOME SEQUENCE ANALYSIS OF THE BIOMASS-DEGRADING FUNGUS

TRICHODERMA REESEI (SYN. HYPOCREA JECORINA) REVEALS

A SURPRISINGLY LIMITED INVENTORY OF

CARBOHYDRATE ACTIVE ENZYMES

Diego Martinez[1‡*#], Randy M. Berka[3‡], Bernard Henrissat[2‡], Markku Saloheimo[4‡], Mikko Arvas[4], Scott E. Baker[8], Jarod Chapman[11], Olga Chertkov[1], Pedro M. Coutinho[2], Dan Cullen[5], Etienne G. J. Danchin[2], Igor V. Grigoriev[11], Paul Harris[3], Melissa Jackson[1], Christian P. Kubicek[9], Cliff S. Han[1], Isaac Ho[11], Luis F. Larrondo[6], Alfredo Lopez de Leon[3], Jon K. Magnuson[8], Sandy Merino[3], Monica Misra[1], Beth Nelson[3], Nicholas Putnam[11], Barbara Robbertse[10], Asaf A. Salamov[11], Monika Schmoll[9], Astrid Terry[11], Nina Thayer[1], Ann Westerholm-Parvinen[4], Conrad L. Schoch[12], Jian Yao[7], Ravi Barbote[1], Mary Anne Nelson[13], Chris Detter[1], David Bruce[1], Cheryl R. Kuske[1], Gary Xie[1], Paul Richardson[11], Daniel S. Rokhsar[11], Susan M. Lucas[11], Edward M. Rubin[11],Nigel Dunn-Coleman[14], Michael Ward[7], Thomas S. Brettin[11]

[1] Los Alamos National Laboratory/Joint Genome Institute, PO Box 1663, Los Alamos, New Mexico 87545 USA. [2]AFMB UMR 6098, CNRS, Universités d'Aix-Marseille I & II, Case 932, 163 Avenue de Luminy, 13288 Marseille, France. [3]Novozymes, Inc., 1445 Drew Ave., Davis, California 95618, USA. [4]VTT Technical Research Centre of Finland, Tietotie 2, Espoo, P.O. Box 1000, 02044 VTT-Espoo, Finland. [5]USDA, Forest Service, Forest Products Laboratory, One Gifford Pinchot Dr., Madison, Wisconsin 53726, USA. [6]Departmento de Genética Molecular y Microbiología. Facultad de Ciericias Biológicas, Pontifica Universidad Católica de Chile and Millennium Institute for Fundamental and Applied Biology, Santiago, Chile. [7]Genencor International, 925 Page Mill Road, Palo Alto, California 94304, USA. [8]Pacific Northwest National Laboratory, P.O. Box 999, Richland, WA 99352, USA. [9]Research Area Gene technology and Applied Biochemistry, Institute of Chemical Engineering, Technische Universitat Wien, Getreidemarkt 9/166, A-1060 Vienna, Austria. [10]Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331. [11]DOE's Joint Genome Institute, 2800 Mitchell Avenue, Walnut Creek, California 94598, USA. [12]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon 93133, USA. [13]Department of Biology, University of New Mexico, Albuquerque, NM 87131, USA. [14]AlerGenetiCa SL, Santa Cruz, Tenerife, Spain.

[‡]These authors contributed equally to this work.

[*]To whom correspondence should be addressed:admar@unm.edu.

[#]Current Address: Department of Biology, University of New Mexico, Albuquerque, NM 87131, USA.

Keywords: cellulases, biomass, biofuels

Abstract

A major thrust of the white biotechnology movement involves the development of enzyme systems which depolymerize biomass to simple sugars which are subsequently converted to sustainable biofuels (e.g., ethanol) and chemical intermediates. The fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*) represents a paradigm for the industrial production of highly efficient cellulases and hemicellulases needed for hydrolysis of biomass polysaccharides. Herein we describe intriguing attributes of the *T. reesei* genome in relation to the future of fuel biotechnology. The *T. reesei* genome sequence was derived using a whole genome shotgun approach combined with finishing work to generate an assembly comprising 89 scaffolds totaling 34 Mbp with few gaps. In total, 9,130 gene models were predicted using a combination of *ab initio* and sequence similarity-based methods and EST data. Considering the industrial utility and effectiveness of its enzymes, the *T. reesei* genome surprisingly encodes the fewest cellulases and hemicellulases of any fungus having the ability to hydrolyze plant cell wall polysaccharides and whose genome has been sequenced. Many genes encoding carbohydrate active enzymes are distributed non-randomly in groups or clusters that interestingly lie between regions of synteny with other Sordariomycetes. Additionally, the *T. reesei* genome contains a multitude of genes encoding biosynthetic pathways for secondary metabolites (possible antibacterial and antifungal compounds) which may promote successful competition and survival in the crowded and competitive soil habitat occupied by *T. reesei*. Our analysis coupled with the availability of genome sequence data provides a roadmap for construction of enhanced *T. reesei* strains for industrial applications.

Trichoderma reesei (teleomorph *Hypocrea jecorina*) is a mesophilic soft-rot ascomycete fungus that is widely used in industry as a source of cellulases and hemicellulases for the hydrolysis of plant cell wall polysaccharides. Discovered as a primary cause of deterioration of military clothing and tents in the South Pacific during World War II, Dr. Elwyn T. Reese and colleagues at the U.S. Army Natick Laboratories isolated a *Trichoderma viride* strain (later renamed *T. reesei strain QM6a* in honor of Dr. Reese) that produced a repertoire of extracellular enzymes that could completely degrade crystalline cellulose[1]. For many years *T. reesei* was believed to be an asexually reproducing fungus; however, application of molecular taxonomic tools identified it to be the anamorph of the pantropical ascomycete *Hypocrea jecorina*[2]. Nevertheless, the organism remains most widely recognized by its former name, *T. reesei*, and has enjoyed a long history of safe use for industrial enzyme production[3] and as a major model system for the study of lignocellulose degradation.

Lignocellulosic biomass from agricultural crop residues, grasses, wood, and municipal solid waste represents an abundant renewable resource that is becoming increasingly important as a future source of biofuels due to environmental and economic issues such as climate change, depletion of fossil fuel reserves and reduce dependence on imported oil. A recent study published by Farrell *et al.*[4] concluded that replacement of gasoline with cellulosic ethanol may substantially reduce greenhouse gases in the atmosphere and decrease global warming. However, a major obstacle that must be overcome to commercialize cellulosic ethanol is the high cost of hydrolyzing biomass polysaccharides to fermentable sugars. Since the cost of cellulase and hemicellulase enzymes accounts for a large fraction of the total price of bioethanol, substantially cheaper sources of these enzymes are needed to foster commercialization of cellulosic ethanol on a wide scale[5]. Consequently, new studies aimed at understanding and improving cellulase efficiency and productivity are at the

forefront of biomass research. Since *T. reesei* represents a paradigm for production of enzymes that hydrolyze biomass polysaccharides, intensive research efforts and considerable government funding have been applied toward the development of better industrial strains that can be implemented for production of fuel ethanol and other biochemicals that are currently derived from non-renewable petroleum-based resources[5]. Fortunately, genetic engineering techniques have been developed to improve industrial enzyme-producing *T. reesei* strains. DNA-mediated transformation systems are available[3], gene knockout protocols have been developed, and there is an active global community of academic researchers. Collectively, these facts have provided the impetus for genome sequencing and analysis with a primary goal of accelerating the world-wide scientific and industrial research on *T. reesei* and the enzymes it produces for conversion of cellulosic biomass.

RESULTS

Features of the *T. reesei* Genome

The genome of *T. reesei* was shotgun sequenced[6] to approximately nine-fold coverage from three libraries with insert sizes of 3 kb, 8 kb and 40 kb totaling 433,863 reads (**Supplementary Table 1**). These data, in addition to more than 6,000 BAC-end sequences[7], yielded a high quality draft assembly using the JGI shotgun assembler JAZZ[8]. A total of 6329 finishing reads were created with custom primers from the 3 kb and 8 kb libraries to close a majority of the gaps and the Phred/Phrap/Consed software package was then used to produce a genome assembly comprising 89 scaffolds and 97 contigs totaling approximately 34 Mb. This figure is only 2.9% larger than the 33 Mb estimate from several karyotyping studies[11,10,9] and agrees with the genome size acquired through physical means. All of the genetic markers that were used in the three studies were also recovered, as well as all protein and RNA sequences in the current release of Genbank (version 161.0). We

are therefore confident that the *T. reesei* genome sequence reported here represents more than 99% of the genome.

Repetitive sequences with similarity to class I and II transposable elements were detected (**Supplementary Note 1**), but all contained multiple stop codons. The apparent absence of active transposons may be explained by active defense mechanisms such as repeat induced point mutation (RIP). These transposable elements totaled less than 1% of the finished genome, which is among the lowest amount found in a fungal genome to date. A repeated hexanucleotide sequence TTAGGG, identical to the telomeric repeat of *Neurospora crassa*, was found at the ends of seven scaffolds in the *T. reesei* genome assembly (**Supplementary Note 1**).

Gene modeling was performed using a combination of homology and *ab initio* methods, selecting a single gene model for each locus (see **Methods**). This yielded 9130 gene models (**Table 1**). This total is relatively close to the number of gene models in *Neurospora crassa*[12], but it is roughly 2500 fewer than the number of predicted genes in *Fusarium graminareum*[13] *(*anamorph*, Gibberella zeae*). The difference in gene number between *F. graminareum* and *T. reesei* is interesting as they share the most recent common ancestor among the genomes listed in **Table 1**[14]. The average gene length in *T. reesei* is 1793 bp with 3.1 exons per gene (average exon length = 508 bp; average intron length = 120 bp). All data, manual cuarations and sequence files are available for viewing and download in the interactive JGI Genome Portal (http://www.jgi.doe.gov/Treesei).

Conserved Synteny in *T. reesei*

To gain insight into the role which environment plays on genome evolution we constructed a comparative map by calculating syntenic regions shared by *T. reesei*, *Fusarium graminearum*[13] and *Neurospora crassa*[12] (**Supplementary Table 2**). As noted in other studies[15], the map in **Figure 1** illustrates segments in which the gene order has changed since the divergence of these species, resulting in large gaps between syntenic blocks. In many cases, these gaps are conserved between *T. reesei* and the other Sordariomycete suggesting that they are prone to frequent insertions, duplications or chromosomal breaks. Regions without synteny to other genomes have been shown to contain genes that are important for the adaptation of the organism[17,15,16]. Another striking feature (Figure 1) is the number of chromosomal rearrangements that have occurred since the divergence of the 3 organisms, clearly illustrating the highly dynamic nature of the genome. The difference in the amount of syntenic coverage of the two other sordariomycetes to *T. reesei* (**Figure 1** and **Supplemental Table 2a)** is consistent with the current understanding about the phylogenetic relationship that *F. graminearum* and *T. reesei* share a more recent common ancestor[18].

To investigate the forces that determine gene synteny, we collected the general features of genes in syntenic blocks vs. gaps (**Supplementary Table 2b**). For many of the metrics there is little difference between the regions, however, we find that there is a large difference in mean exon size (89 nucleotides, p-value of 2.2 e-16) in the *F. graminearum* blocks and gaps comparison. In an analysis of Interpro[19] domain content between the two groups of genes (**Supplementary Table 2c**) it is evident that there is a noticeable difference in the number of genes with the domain IPR001680 (G-protein beta WD-40 repeat). Genes with this domain appear to have unusually large exons, and can account for some of the shift in

16

mean exon size. Another interesting finding from the InterPro comparison in

Supplementary Table 4c is that the InterPro domain IPR000254 (Cellulose-binding region, fungal) is found only in genes that lie in syntenic gaps, and is not found in *T. reesei* genes that are in syntenic blocks.

Protein Domains in *T.reesei*

We compared the protein domains encoded by the *T. reesei* genome to those of 13 fungal genomes using InterProScan[19] to find regions in proteins with known functions. Compared to sequenced species within Pezizomycotina, the proteome of *T. reesei* is under-represented in many proteins with known functions (**Supplementary Table 3**). In particular, *T. reesei* differs in its content of proteins related to plant biomass degradation (see CAZymes section below). In line with its natural role as a necrophyte, *T. reesei* lacks several families related to infecting and degrading living plant tissue, such as pectate lyases and pectinesterases. In addition, no tannase and feruloyl esterase family members were found, leaving *T. reesei* apparently handicapped in the degradation of hemicellulose.

Carbohydrate Active enZymes (CAZymes) in *T. reesei* and Comparative Analysis with Other Fungi

Carbohydrate-active enzymes or CAZymes are categorized into different classes and families in the CAZy database (http://www.cazy.org)[20]. CAZymes that cleave, build and rearrange oligo- and polysaccharides play a central role in the biology of *T. reesei* and other fungi. In addition, the CAZyme family is key to optimizing the ability of *T. reesei* to degrade biomass. Given the relative importance of this gene family to the biotechnology community, we performed a detailed examination of the CAZome of *T. reesei* and compared it with the

corresponding gene subsets from 13 fungi whose genomes have been sequenced (**Table 2**).

One might expect that *T. reesei*, which is known as an efficient plant polysaccharide degrader and has been an important model of the degradation system, would contain expansions of genes involved in breakdown of these cell wall compounds. Contrary to this expectation, *T. reesei* is surprisingly poor in genes encoding glycoside hydrolases (GHs). With a total of 200 GH-encoding genes, it ranks only penultimate of the Sordariomycetes in our dataset (**Table 1**). This figure is also slightly below the average number of GHs found in this lineage (211), though the difference does not exceed the standard deviation (SD=32). The high standard deviation among Sordariomycetes may reflect adaptations of GHs to the ecology of these species: the two phytopathogens *M. grisea* and *F. graminearum* are the richest. Compared to other fungal lineages, Sordariomycetes represent the second most GH-rich lineage, preceded only by Eurotiomycetes which have an average number of 265 GHs and a more homogeneous GH distribution (SD=19).

With 103 glycosyltransferases (GTs), *T. reesei*, is close to the average among Sordariomycetes (96) (**Table 2**). This enzyme class has less variability in Sordariomycetes than GHs (SD=15), as is also noticeable in the other phyla in our dataset. This trend is maintained both for intra-lineage variability and inter-lineage variability suggesting that GTs possess basal intracellular activities, and variations in composition may reflect species drift rather than environmental pressure.

The enzymes involved in plant polysaccharide depolymerization frequently carry a carbohydrate-binding module (CBM) appended to the catalytic domain. Unexpectedly, the *T.*

18

reesei genome has the smallest number of CBM-containing proteins among the Sordariomycetes in our dataset (**Table 2**). However, it should be noted that the high number of CBMs in Sordariomycetes (the highest in this dataset) is essentially due to a significant enrichment in the two phytopathogenic fungi, *F. graminearum* and *M. grisea*. Similarly, *T. reesei* has the lowest number (16) of carbohydrate esterases (CEs) compared to other Sordariomycetes. The difference to the average among Sordariomycetes (32) is approximately equal to the standard-deviation (SD=15).

The Sordariomycetes, including *T. reesei*, show a relative paucity in polysaccharide-lyases (PLs), a category that typically contains three to four genes, except for *F. graminearum* with an expansion of 20 genes. Such a high number of PLs is only found in the Eurotiomycetes with an average number of 18 PLs. No PLs are found in unicellular Ascomycetes. In conclusion, the *T. reesei* genome encodes a number of CAZymes that is slightly below the average found among Sordariomycetes. Detailed statistical analyses are presented in **Supplementary Note 2** and **Supplementary Table 4**.

Surprisingly, a thorough inspection of the *T. reesei* genome revealed only seven genes encoding well-known cellulases (endoglucanases and cellobiohydrolases), which places *T. reesei* at the bottom of the list of fungi able to degrade the plant cell walls (**Table 3**). This trend is further amplified if one adds the family GH61 proteins (**Table 3**). Hemicellulose comprises a diverse group of complex polysaccharides and their complete degradation requires an arsenal of enzymes. With only 16 genes, *T. reesei* has the smallest set of hemicellulases among all fungi analysed (**Table 4**). Similarly, *T. reesei* has the smallest set of enzymes for the breakdown of pectin among the plant cell wall degrading fungi (**Supplementary Table 5**).

Protein Secretion

T. reesei is an extraordinarily efficient producer of extracellular enzymes, and industrial strains can achieve production levels of 100 g/L of extracellular protein[21]□. This remarkable capability suggests that the protein secretion machinery of *T. reesei* is exceptionally efficient. Consequently, the content of genes encoding secretory pathway components in its genome is of particular interest. Not surprisingly, homologues of proteins that function in the secretory pathway of *S. cerevisiae* were found in the *T. reesei* genome. While they are generally present as single-copy genes in *T. reesei* and show greater similarity to the yeast orthologs than to their mammalian counterparts, there are a few notable exceptions to this trend. *T. reesei* appears to have three proteins whose closest homologue in yeast is protein disulphide isomerase, Pdi1p. This could be connected to the fact that the major secreted cellulases of this fungus have many disulphide bonds[22]. The ER-associated protein degradation (ERAD) pathway of *T. reesei* appears to be more redundant than the secretory pathway in general, since two orthologs of the yeast *DER1* and *UFD1* genes are found. We found clear homologues of most of the other known ERAD components in *T. reesei*, even though Pel *et al.*[23] reported a lack of orthologs or little sequence similarity to yeast ERAD components in the *A. niger* genome.

Orthologs of most *S. cerevisiae* proteins that are known to be involved in protein trafficking can be found as single copies in the *T. reesei* genome. Whereas yeast lacks counterparts of the mammalian GTPase proteins Rab2, Rab4 and Rab5 and Arf6 and Arf10, *T. reesei* as well as *N. crassa* appear to have orthologs of these signalling proteins involved in membrane fusion or vesicle budding in diverse cellular locations (**Supplementary Table 6**). It is also of interest that the t-SNARE protein Sso1p of yeast, a receptor for the secretory

vesicles on the plasma membrane, has two homologues in *T. reesei*, and a recent study indicates that the two *SSO1* homologues have divergent functions[24]□. Taken together, these findings suggest that the membrane trafficking system in *T. reesei* is more diverse than in *S. cerevisiae*.

CAZyme "gene clusters" in *T. reesei*

In the *T. reesei* genome, there is non-random organization of genes that encode a portion of the CAZymes. In a previous study nine known genes involved in cellulose and hemi-cellulose degradation were shown to be colocated in several areas of the genome[7]. We have extended this work to the location of all CAZymes in the genome and found that in total, 130 of the 316 (41%) CAZyme genes are found in 26 discrete regions ranging from 14 kb to 275 kb in length (roughly 2.4 Mb or 7% of the genome) (**Figure 2** and **Supplementary Table 7**). These regions contain an average five-fold increase in CAZyme gene density compared to the expected density for randomly distributed genes, and based on the hypergeometric distribution (see Methods) we have calculated the p-value of the clusters, which ranges from .015 to 1e-4. Each region contains from two (as adjacent pairs) to as many as ten CAZyme genes.

To gain insight into how such clusters arise, we analyzed the number of orthologs within the clusters. We find that 95 of the 130 (73%) CAZy genes that are in clusters are in gaps of synteny. We also observe that of those 95 CAZy genes, 69 are orthologs with *F. graminearum* (72%). In addition, there are a mere 16 CAZyme orthologs that are in synteny (with *F. graminearum*), indicating that gene movement is the major factor in the organization of the clusters, while gene duplications play a minor role. With respect to the non-orthologous CAZy genes (the potential duplicates), all have homologs in almost all the

fungal genomes sequenced to date. In addition, few CAZys in the same cluster are from the same CAZyme family, with a few notable exceptions (see Supplementary Table 7); only 10 genes in 4 different clusters are from the same sub family, including a pair of GH3s and a triplet of GH3s. This leads us to the conclusion that the few paralogs that are co-located in the clusters indicate that gene relocation rather than duplication are responsible for the formation of the CAZy clusters.

The profile of CAZymes found in the clusters suggests a specific biological role. Approximately 70% of the CAZyme genes in the clusters are GHs. The finding that 24% of the GTs and 46% of the GHs in the genome are found in these regions indicates that the majority of the CAZymes in these clusters are involved in polysaccharide degradation (**Supplementary Table 7**). This is supported by the finding that many of the genes previously shown to be involved in plant cell wall degradation fall into the CAZyme rich regions (**Supplementary Table 8**). Three of the four expansin-like genes in *T. reesei*, including the previously described swollenin gene[25] are located in these clusters (**Figure 2**). It is intriguing that the few GTs found in CAZyme clusters are largely mannosyltransferases, chitin synthases (four of nine in *T. reesei*), α-glycosyltransferases and β-glycosyltransferases, enzymes which may be involved in synthesis of fungal cell walls[26].

A portion of the data from two transcriptomics projects identifying *T. reesei* genes induced by sophorose[27] and cellulose[28] were mapped to the genome. While not all of the clustered GHs were co-expressed in the above studies, we found four examples in which adjacent or nearly adjacent genes were co-expressed (**Figure 2**), giving further evidence for the biological importance of the CAZYme clustering. Interestingly, in these regions there is no syntenic signal with any of the other fungal genomes as shown in **Figure 1**, suggesting that

these genes are reordered in *T. reesei,* and that this organization is an evolutionary advantage for the fungus.

Several of the regions of high CAZyme gene density also contain genes encoding proteins involved in secondary metabolism (**Supplementary Table 7**). Specifically, five of the 25 CAZyme clusters contain either a polyketide synthase (PKS) gene or a non-ribosomal peptide synthase (NRPS) gene. In particular, we found two non-reducing PKS genes (scaffold_1:410000-530000 and scaffold_6:10000-148000) that in our phylogenetic analysis (maximum likelihood performed with PHYML and RAXYML, data not shown) appear in a clade with previously undescribed PKS genes. In addition, the PKS gene in the region of scaffold_6:10000-148000 is fused with an NRPS gene that resides in a clade with NRPS genes involved in lovastatin and citrinin production (maximum likelihood performed with PHYML and RAXML, data not shown). Another intriguing finding is that *T. reesei* has retained most NRPS paralogues as compared to other Sordariomycetes analyzed thus far. **Supplementary Table 9** lists the NRPS and PKS genes found in the *T. reesei* genome.

DISCUSSION

Remarkably, *T. reesei* has the smallest repertoire of genes encoding the three categories of enzymes involved in depolymerization of plant cell wall polysaccharides: cellulases, hemicellulases and pectinases (**Tables 3, 4** and **Supplementary Table 5,** respectively). This is unexpected since the cellulolytic enzyme machinery of *T. reesei* is efficient and represents the paradigm for the enzymatic breakdown of cellulose and hemicellulose. An inability to rationalize the diversity observed in the composition of cellulolytic enzymes among fungal proteomes suggests that a lack of understanding persists regarding plant cell wall degradation. Thus, there may be room for improvement of *T. reesei* strains by

augmenting its inventory with genes/enzymes from other sources. On the other hand, *T. reesei* successfully competes in nature as a degrader of cellulose and hemicellulose, and its limited enzyme set is sufficient for utilization of these substrates. To what degree its success is enhanced by an array of secondary metabolites is unknown. However, it is tempting to speculate that the clustering the GH genes (in some cases near secondary metabolite genes) has enabled *T. reesei* to control the expression of these genes more efficiently.

The *T. reesei* genome revealed that several enzyme families involved in polysaccharide degradation are reduced or absent. Of all the possible CAZyme genes involved in pectin degradation, only members of GH28 are found, and there is no expansion of this family that could compensate for the lack of other pectinolytic enzymes. The deficiency of pectinolytic enzymes is confounding when comparing *T. reesei* with other Sordariomycetes (**Supplementary Table 2**), but it is consistent with the poor growth of the species on D-galacturonic acid and L-rhamnose[29], two constituents of the pectin backbone. L-Arabinose and D-galactose, which make up the majority of the side chains in "hairy regions" of pectin, are readily metabolized. One possible explanation is that the pectin backbone is depolymerized by other fungi and bacteria in the soil environment where *T. reesei* exists primarily as a secondary colonizer. The absence of invertase (EC 3.2.1.26; family GH32) is also consistent with the fungus being a secondary colonizer since sucrose is probably consumed rapidly by primary colonizers.

Previous studies indicate that the locations of genes in both bacterial and eukaryotic genomes are not necessarily random[30]. In fungi, there are examples of gene clusters that are involved in the production of secondary metabolites, including NRPS/PKS pathways, or oxidation of substrates, *e.g.* cytochrome P450 genes in *Phanerochaete chrysosporium*[31]. In

several *Clostridium* species of bacteria there is an intriguing parallel to the *T. reesei* CAZyme clusters in that the genes of the cellulosome complex encoding GH enzymes needed for cellulose and hemicellulose degradation are also clustered[32]. However, the distances are much shorter between GH genes than in *T. reesei*, excluding the cases shown in **Figure 2**. Thus, in *Clostridium* cellulosome gene clusters as well as in *the T. reesei* CAZyme clusters functional coupling of genes involved in the hydrolysis of cellulose and hemicellulose creates pressure to maintain their position relative to one another. This is in agreement with the chemical complexity of plant cell wall polysaccharides which requires a diverse mixture of enzymes for complete depolymerization. Given these observations it is reasonable to conclude that the clustering of the CAZY genes is favored by selection favoring enhanced degradative efficiency and coordinated regulation that a co-localization strategy may offer.

The concentration of CAZyme genes (primarily GHs) in syntenic gaps with any other species tested further supports the notion that selective pressure can maintain the clustering of genes involved in biomass degradation. In comparison, previous studies[15-17] indicate that syntenic gaps in other genomes are enriched in genes that are important for species-specific attributes. While it is possible that duplications may play a role in the loss of synteny, the CAZyme clusters in *T. reesei* show little evidence of expansion in comparison with the other fungi analyzed. Indeed, there are few clusters that contain appreciable numbers of genes from the same subfamily (**Supplementary Table 7)** indicating that recent duplication has not played an important role. It is therefore likely that the majority of the breaks in synteny where CAZYme genes are clustered arise from movement of CAZyme genes into these regions, followed by pressure to fix the genomic rearrangements in the population.

The reduction in duplicated genes in *T. reesei* could be attributed to the effects of RIP (repeat induced point mutation), similar to the limitation seen in *N. crassa (**Supplementary Note 1**).* As previously mentioned, a RIP-like mutation pattern is observed in the transposons of *T. reesei*, however, the density of mutations is lower than in *N. crassa*. This could explain why the genome size of *T. reesei* and *N. crassa* are similar and contain few intact repeats, and it may be a reason for the lack of gene family expansion in GHs, forcing the organism to favor gene translocations in the pursuit of environmental adaptation.

The biased placement of several secondary metabolism genes near CAZyme clusters presents an intriguing possibilty - that *T. reesei* in the process of acquiring nutrients must fend off competitors. In addition, the number of conserved PKS and NRPS genes in *T. reesei* suggests that survival requires an arsenal of antimicrobial secondary metabolites, particularly in light of the limited repertoire of CAZymes. The only GH family that contains any appreciable enrichment is that of the chitinase genes in family GH18 (**Table 3**), nearly half of which can be found in clusters. Other members of genus *Trichoderma* (*e.g. T. harzianum, T. atroviride*) are capable of mycoparasitism, and both chitinases and secondary metabolites could be important in attacking other fungi[33].

Although the organization of GH genes may contribute to the ability of *T. reesei* to efficiently degrade plant material, the lack of key enzyme activities clearly suggests opportunities for industry to generate improved enzyme cocktails that may be applied for the conversion of plant biomass to fermentable sugars. Since complete hydrolysis of cellulosic and hemicellulosic substrates requires multiple enzymes acting synergistically, development of superior enzyme blends will likely occur via genetic engineering of suitable industrial strains.

The capacity for secreting copious amounts of extracellular enzymes, availability of genetic tools and straightforward, inexpensive fermentation of *T. reesei* make it an ideal candidate for production of enzymes useful for conversion of biomass feedstocks such as corn stover, cereal straw, and switch grass[34] to fuel ethanol and other chemicals that are currently derived from non-renewable resources. Production of these enzymes at economically viable levels will require an increased understanding of the dynamics of cell growth and enzyme production. Mathematical and kinetic models are being developed to optimize these processes[35], and the availability of a complete genome sequence will provide a blueprint to improve the models and to empower strain improvement strategies for creating superior enzyme mixtures from a single, highly productive strain.

METHODS

Automated Annotation

In addition to the methods described in Grigoriev *et al.*[36] genes in the *T. reesei* genome were predicted using an *ab initio* gene predictor, Fgenesh[37], specifically trained for this genome, and two homology-based gene predictors, Fgenesh+ (www.softberry.com) and Genewise[38]. All three methods predict only CDS regions in genes, which we then corrected and where possible extended into putatively full-length genes using 42,916 *T. reesei* ESTs. Finally, using a heuristic approach implemented in the JGI pipeline, we combined all predicted gene models to produce a non-redundant set of genes, in which a single best gene model per locus was selected on basis of sequence similarity to known proteins and support by available ESTs. This representative set included 9,130 genes and was subject to manual curation and genome analysis described in this work.

The majority (82%) of predicted genes contain multiple exons with average of 3.1 exons per gene. Average gene density, similar between most of the larger scaffolds, is 3.7 kb/gene. Average gene, transcript and protein lengths are 1.8 Kb, 1.6 kb and 492 amino acids, respectively (**Table 1**). In total 7,887 (86%) gene models were predicted to be complete. There are 42,916 ESTs that support 46.1% of the predicted genes. Approximately 94% of the predicted proteins show sequence similarity to other proteins, primarily from fungi. A total of 2,164 manually curated genes from version 1.2 of *T. reesei* Genome Portal were mapped forward to version 2.0.

Genes were annotated and classified according to Gene Ontology (GO)[39], eukaryotic orthologous groups (KOGs)[40], and KEGG metabolic pathways[41]. We assigned GO terms to 4,977 (54.5%) of the predicted *T. reesei* proteins including 3,547, 1,913 and 4,651 genes with molecular function, cellular component and biological process, respectively. We also assigned 5,420 (59.4%) proteins to KOG clusters. We assigned 751 distinct EC numbers to 2,264 (25%) proteins mapped to KEGG pathways.

Manual Curation

Gene function assignments were manually curated for 2,164 gene models using an interactive website (http://genome.jgi.doe.gov/Trire2). To assign confidence to these functional calls as well as to standardize the nomenclature methods, a qualifier system was employed based on the homolog for which a functional assignment was made, and is used throughout the paper. This nomenclature was based on the following naming convention: A three letter code was assigned to a gene only when the gene had experimental evidence in *T. reesei*. If an experiment had not been performed in *T. reesei*, the tag tre<gene_id> was used, for example, tre167435. The definition line, or "def_line" was assigned based on

sequence similarity to proteins in other organisms and Interpro domains. If the best sequence similarity was above 80% identity and 80% coverage (calculated as alignment length divided by predicted protein length) to a protein that had experimental evidence in publication, no def_line qualifier was used. If the sequence similarity was above 70% id and 70% coverage, yet lower than 80% id and 80% coverage to a protein that was described in publication, the def_line qualifier "Candidate" was used (for example candidate α,α-trehalase). If a homolog of a protein described in publication was above 50% id and 50% coverage, the def_line qualifier "Related to" was used, as in "Related to α-fucosidase." Below this last threshold, all def_lines are tagged with the qualifier "Hypothetical." Unknown and hypothetical proteins with hits to only other unknown hypothetical proteins are assigned the def_line "Conserved Hypothetical."

Calculation of Syntenic Blocks

The areas of relationship known as syntenic (meaning *same ribbon*) regions or syntenic blocks are anchored with orthologs (calculated as mutual best hits or bi-directional best hits) between the two genomes in question, and are built by controlling for the minimum number of genes, minimum density, and maximum gap (genes not from same genome area) as compared with randomized data as described in Sankoff et al.[42] A version of the algorithm was programmed in PERL, and runs in less than one minute on an AMD Opteron dual CPU machine with 6 Gigabytes of RAM. This savings in time is largely due to the requirement the orthologs be precalculated from BLAST results (min e-value e-5, 40% coverage required). Code available upon request.

While this technique may cause artificial breaks, it highlights regions that are dynamic and picking up a large number of insertions or duplications. From the analysis shown in

Supplementary Table 2 *N. crassa* shares 5624 mutual best hit (MBH) genes with *T. reesei* (62% of *T. reesei* genes) and *F. graminearum* shares 6580 MBH genes with *T. reesei* (72% of *T. reesei* genes) that have maintained their general location since divergence from their most recent common ancestor.

Protein Domains

The proteomes used in this study included *Aspergillus fumigatus*, *A. nidulans*, *F. graminearum* (not yet published), *T. reesei*, *M. griseae*, *N. crassa*, *Ashbya gossypii*, *Candida albicans*, *C. glabrata*, *Debaryomyces hansenii*, *Kluyveromyces lactis*, *Yarrowia lipolytica* and *S. cerevisiae.* The number of genes found to have a certain Interpro entry was counted. In order to get robust results that would not be clouded by differences in sequencing coverage, assembly or version of the genomes used, we searched for over-represented Interpro entries by selecting those which had at least twice as many corresponding genes in *T. reesei* than in any other euascomycete and *vice-versa* for under-represented entries. Differences in Interpro entry counts can be due to actual presence or absence of a domain or mutations in the domain's sequence that renders it unrecognisable to InterProScan. To classify the results accordingly and verify them we carried out BLAST searches and studied alignments of homologous genes.

Detection of carbohydrate-active enzymes in fungal proteomes

The search for carbohydrate-active modules (GHs, GTs, PLs and CEs) and their associated carbohydrate-binding modules (CBMs) in *T. reesei* was performed exactly as for the daily updates of the Carbohydrate-Active enZYme (CAZy) database (http://afmb.cnrs-mrs.fr/CAZY/). Briefly, the sequences of the proteins in CAZy were cut into their constitutive modules (catalytic modules, CBMs and other non-catalytic modules or domains of unknown

function). The resulting fragments were assembled and formatted as a sequence library for BLAST searches. Accordingly, each protein model from *T. reesei* (and other fungal proteomes) was BLASTed against the library of approximately 100,000 individual modules using a database size parameter identical to that of the NCBI non-redundant database. All models that gave an e-value better than 0.1 were automatically kept and manually analyzed. Manual analysis involved examination of the alignment of the model with the various members of each family (whether of catalytic or non-catalytic modules), with a search of the conserved signatures/motifs characteristic of each family. The presence of the catalytic machinery was verified for borderline cases whenever known in the family. The models that displayed the usual features that would lead to their inclusion in the CAZy database were kept for annotation and classified in the appropriate class and family.

Functional Annotation of Protein Models Corresponding to Carbohydrate-active Enzymes

The analysis of the sequence-based families of GHs and GTs shows that those families rarely coincide with a single substrate (or product) specificity[43]. As a consequence, many of these families group together enzymes that have different EC numbers. Our annotation strategy aims at producing (as much as possible) annotations that will "age" well, *e.g.*, that are designed to survive experimental validation while avoiding over-interpretation. For instance, in a family that contains β-mannosidases, β-galactosidases and β-glucuronidases, all enzymes hydrolyze equatorially oriented glycosidic bonds. A strong similarity to β-galactosidases allows annotation as "candidate β-galactosidase", but if similarity is not sufficient for a safe prediction of substrate specificity, the best possible annotation is "candidate β-glycosidase". Each protein model kept from the modular annotation step was thus annotated using that scheme. The proteins were BLASTed again against the manually

curated CAZy database, and we assigned a functional annotation according to the relevance of the BLAST matches. Only in the cases where the enzyme of the species itself has been experimentally characterized was the protein given an EC number. All uncharacterized protein models were thus at best "candidates" or "related to" or "distantly related to" their characterized match. Because the threshold of similarity that correlates with a change of substrate specificity is extremely variable from one family to another, the criteria were tightened or loosened appropriately for several families.

Fungal Cazome Comparisons

We utilized several statistical analyses to identify the significant features in the comparison of the sets of carbohydrate-active enzymes encoded by 13 fungal genomes taking into account both taxonomic and CAZyme families variability. Basically, the approach consisted in applying a *chi*-square independence test and other statistical analyses to identify the most unexpected points for a given CAZyme family / species according to the general distribution.

For each class of CAZymes, the statistical test first required placing the data in a table of *k* columns (representing the different families) and *l* rows (representing the different species). The A_{ij} value will represent the number of CAZymes from family *i* and species *j*. We next calculate the values of:

$$\blacksquare_{ij} = \frac{\sum A_{ik} \sum A_{lj}}{\sum A_{kl}}$$

then,

$$\sum \frac{\left| A_{ij} - \blacksquare_{ij} \right|^2}{\blacksquare_{ij}}$$

32

The last value allows the rejection of the *chi*-square independence hypothesis, and the A_{ij} that contribute the more to the total sum represents the points (families) that are the most significantly different for a given species.

Gene Cluster Identification

The gene families in question were collected by visual inspection using the JGI Genome Portal for the *T. reesei* genome. A cluster is defined as a region containing a statistically higher proportion of a particular gene family and must begin and end with a gene from the family in question. We then calculated the probability that a proportion in the cluster of the particular gene family is higher than the current one using the Hypergeometric distribution (expressed as a p-value). In gathering such clusters, it is possible to take a smaller section and get a higher p-value, however, our goal was to take the longest reasonable cluster that had a p-value less than .05 (outside the 95% confidence interval). In the CAZyme clusters presented here the mean p-value is 3.9 e-3, and only 4 of the 25 clusters has a p-value > .01, outside the 99% interval, but still less than .05.

Accession Number

The *T. reesei* nucleotide sequence and annotation data have been deposited in Genbank under accession number AAIL00000000.

Note: Supplementary information is available on the Nature Biotechnology website.

Acknowledgements

Meirelles for their consultation on statistics. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, Los Alamos National Laboratory under contract No. W-7405-ENG-36 and NIH grant GM060201. The work was also funded in part by the European Commission (STREP FungWall grant, contract: LSHB - CT- 2004 - 511952).

Supplementary information is available on-line at

http://www.nature.com.libproxy.unm.edu/nbt/journal/v26/n5/suppinfo/nbt1403_S1.html

Author Contributions

Statement of Work: DM CAZyme gene organization and synteny calculation, RB annotation of transcription machinery, BH, EGJD and PMC annotation and comparative analysis of CAZyme genes, MA protein domains, MS and MW annotation of the secretory pathway, SB and BR annotation of secondary metabolism, DC annotation of repetitive elements and RIP, OC, CH and TB genome finishing, IG and AS gene modeling, JG, IH and NP genome assembly, PH annotation of DNA replication, repair and recombination, CK and MS annotation of signaling pathway, LL annotation of oxidases, AL annotation of transcription factors, JM central metabolism, SM annotation of sexual development, BN amino acid metabolism, AT, NT, GX, MJ and MM gene model manual curation, RB annotation of transporters, AWP annotation of microtubules and molecular motors, JY annotation of proteases, PR and SL sequencing.

Competing Interests Statement

The authors declare that they have no competing financial interests.

References

1. Mary Mandels & Elwin T. Reese Induction of cellulase in Trichoderma viride as influnced bycarbon sources and metals. *Journal of Bacteriology* **73**, 279-283 (1957).

2. Kuhls, K. et al. Molecular evidence that the asexual industrial fungus Trichoderma reesei is a clonal derivative of the ascomycete Hypocrea jecorina. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 7755-7760 (1996).

3. NEVALAINEN, H., SUOMINEN, P. & TAIMISTO, K. ON THE SAFETY OF TRICHODERMA-REESEI. *Journal of Biotechnology* **37**, 193-200 (1994).

4. Farrell, A. et al. Ethanol can contribute to energy and environmental goals. *Science* **311**, 506-508 (2006).

5. Patel-Predd, P. Overcoming the hurdles to producing ethanol from cellulose.*Environmental Science & Technology* **40**, 4052-4053 (2006).

6. Detter, J. et al. Isothermal strand-displacement amplification applications for high-throughput genomics. *Genomics* **80**, 691-698 (2002).

7. Diener, S. et al. Insight into Trichoderma reesei's genome content, organization and evolution revealed through BAC library characterization. *Fungal Genetics and Biology* **41**, 1077-1087 (2004).

8. Shapiro, H. Outline of the assembly process: Jazz, the JGI in-house assembler. (2005).

9. HERRERAESTRELLA, A. et al. ELECTROPHORETIC KARYOTYPE AND GENE ASSIGNMENT TO RESOLVED CHROMOSOMES OF TRICHODERMA SPP. *Molecular Microbiology* **7**, 515-521 (1993).

10. CARTER, G. et al. CHROMOSOMAL AND GENETIC-ANALYSIS OF THE ELECTROPHORETIC KARYOTYPE OF TRICHODERMA-REESEI - MAPPING OF THE CELLULASE AND XYLANASE GENES. *Molecular Microbiology* **6**, 2167-2174 (1992).

11. MANTYLA, A. et al. ELECTROPHORETIC KARYOTYPING OF WILD-TYPE AND MUTANT TRICHODERMA-LONGIBRACHIATUM (REESEI) STRAINS. *Current Genetics* **21**, 471-477 (1992).

12. Galagan, J. et al. The genome sequence of the filamentous fungus Neurospora crassa. *Nature* **422**, 859-868 (2003).

13. Cuomo, C. et al. The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. *Science* **317**, 1400-1402 (2007).

14. Taylor, J. & Berbee, M. Dating divergences in the Fungal Tree of Life: review and new analyses. *Mycologia* **98**, 838-849 (2006).

15. Galagan, J. et al. Sequencing of Aspergillus nidulans and comparative analysis with A-fumigatus and A-oryzae. *Nature* **438**, 1105-1115 (2005).

16. Machida, M. et al. Genome sequencing and analysis of Aspergillus oryzae. *Nature* **438**, 1157-1161 (2005).

17. Nierman, W. et al. Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. *Nature* **438**, 1151-1156 (2005).

18. Zhang, N. et al. An overview of the systematics of the Sordariomycetes based on a four-gene phylogeny. *Mycologia* **98**, 1076-1087 (2006).

19. Zdobnov, E. & Apweiler, R. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-848 (2001).

20. Coutinho, P. & Henrissat, B. Cabohydrate-active enzymes: An integrated database approach. *Recent Advances in Carbohydrate Bioengineering* 3-14 (1999).

21. Cherry, J. & Fidantsef, A. Directed evolution of industrial enzymes: an update. *Current Opinion in Biotechnology* **14**, 438-443 (2003).

22. DIVNE, C. et al. THE 3-DIMENSIONAL CRYSTAL-STRUCTURE OF THE CATALYTIC CORE OF CELLOBIOHYDROLASE-I FROM TRICHODERMA-REESEI. *Science* **265**, 524-528 (1994).

23. Pel, H. et al. Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88. *Nature Biotechnology* **25**, 221-231 (2007).

24. Valkonen, M. et al. Spatially segregated SNARE protein interactions in living fungal cells. *Journal of Biological Chemistry* **282**, 22775-22785 (2007).

25. Saloheimo, M. et al. Swollenin, a Trichoderma reesei protein with sequence similarity to the plant expansins, exhibits disruption activity on cellulosic materials. *European Journal of Biochemistry* **269**, 4202-4211 (2002).

26. Cabib, E. et al. The yeast cell wall and septum as paradigms of cell growth and morphogenesis. *Journal of Biological Chemistry* **276**, 19679-19682 (2001).

27. Foreman, P. et al. Transcriptional regulation of biomass-degrading enzymes in the filamentous fungus Trichoderma reesei. *Journal of Biological Chemistry* **278**, 31988-31997 (2003).

28. Arora, D.K. & Berka, R. Genes and Genomics, Volume 5. 444 (2005).

29. Druzhinina, I. et al. Global carbon utilization profiles of wild-type, mutant, and transformant strains of Hypocrea jecorina. *Applied and Environmental Microbiology* **72**, 2126-2133 (2006).

30. Hurst, L., Pal, C. & Lercher, M. The evolutionary dynamics of eukaryotic gene order. *Nature Reviews Genetics* **5**, 299-310 (2004).

31. Martinez, D. et al. Genome sequence of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78. *Nature Biotechnology* **22**, 695-700 (2004).

32. Doi, R. & Kosugi, A. Cellulosomes: Plant-cell-wall-degrading enzyme complexes. *Nature Reviews Microbiology* **2**, 541-551 (2004).

33. Seidl, V. et al. A complete survey of Trichoderma chitinases reveals three distinct subgroups of family 18 chitinases. *Febs Journal* **272**, 5923-5939 (2005).

34. Rosgaard, L. et al. Efficiency of new fungal cellulase systems in boosting enzymatic degradation of barley straw lignocellulose. *Biotechnology Progress* **22**, 493-498 (2006).

35. Tholudur, A., Ramirez, W. & McMillan, J. Mathematical modeling and optimization of cellulase protein production using Trichoderma reesei RL-P37. *Biotechnology and Bioengineering* **66**, 1-16 (1999).

36 Arora, D.K., Berka, R. & Singh, G.B. Bioinformatics, Volume 6. 350 (2006).

37. Salamov, A. & Solovyev, V. Ab initio gene finding in Drosophila genomic DNA. *Genome Research* **10**, 516-522 (2000).

38. Birney, E. & Durbin, R. Using GeneWise in the Drosophila annotation experiment. *Genome Research* **10**, 547-548 (2000).

39. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29 (2000).

40. Koonin, E. et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology* **5**, (2004).

41. Kanehisa, M. et al. The KEGG resource for deciphering the genome. *Nucleic Acids Research* **32**, D277-D280 (2004).

42. Hoberman, R., Sankoff, D. & Durand, D. The statistical significance of max-gap clusters. *Comparative Genomics* **3388**, 55-71 (2005).

43. Stam, M. et al. Evolutionary and mechanistic relationships between glycosidases acting on alpha- and beta-bonds. *Carbohydrate Research* **340**, 2728-2734 (2005).

44. Dean, R. et al. The genome sequence of the rice blast fungus Magnaporthe grisea. *Nature* **434**, 980-986 (2005).

45. Jones, T. et al. The diploid genome sequence of Candida albicans. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 7329-7334 (2004).

46. Goffeau, A. et al. Life with 6000 genes. *Science* **274**, 546-& (1996).

47. Dujon, B. et al. Genome evolution in yeasts. *Nature* **430**, 35-44 (2004).

48. Wood, V. et al. The genome sequence of Schizosaccharomyces pombe. *Nature* **415**, 871-880 (2002).

49. Loftus, B. et al. The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. *Science* **307**, 1321-1324 (2005).

Table 1. General features of fungal genomes compared to *T. reesei*.

Organism	Size	No. Genes	% Coding	%GC
T. reesei	33.9Mb	9130	40.40%	52.0%
F. graminearum	36.1 Mb	11640	56.24%	48.3%
N. crassa	38.7 Mb	10620	38.50%	49.6%
M. grisea	39.4 Mb	12841	50.40%	52.0%
A. nidulans	30.1 Mb	10701	58.80%	50.3%
S. cerevisiae	12.0 Mb	5885	72.55%	38.3%
P. chrysosporium	34.5 Mb	10048	42.22%	56.8%

Table 2. Total number of CAZyme families, by class, in the 13 fungal genomes analyzed. *T. reesei* appears in bold. Averages are given per taxonomic group. Abbreviations: GH, glycoside hydrolase; GT, glycosyltransferase; CBM, carbohydrate-binding module; CE, carbohydrate esterase; PL, polysaccharide lyase. The highest and lowest number of CAZyme entries in each enzyme class is indicated in red and blue, respectively. Species abbreviations and genome references, A.nid: *Aspergillus nidulans* strain FGSC A4 [15], A.fum: *Aspergillus fumigatus* clinical isolate Af293[17], A.ory: *Aspergillus oryzae* strain RIB40[16], M.gris: *Magnaporhe grisea* strain 70-15[44], N.cra: *Neurospora crassa* strain 74A[12], T.ree: *Trichoderma reesei* (current paper), F.gra: *Fusarium graminearum* strain PH-1[13], C.alb: *Candida albicans* strain SC5314[45], S.cer: *Saccharomyces cerevisiae* strain S288C [46], C.glab: *Candida glabrata* strain CBS138[47], S.pom: *Schizosaccharomyces pombe* strain 972H[48], C.neo: *Cryptococcus neoformans* strain JEC21[49], P.chr: *Phanerochaete chrysosporium* strain RP-78[31].

Lineages		Species	GH	Avr. GH	GT	Avr. GT	CBM	Avr. CBM	CE	Avr. CE	PL	Avr. PL
Ascomycetes	Eurotio.	A.nid	247		91		36		29		19	
		A.fum	263	265	103	103	55	40	29	28	13	18
		A.ory	285		114		30		26		21	
	Sordario.	M.gris	231		94		58		47		4	
		N.cra	171	211	76	96	39	49	21	32	3	8
		T.ree	200		103		36		16		3	
		F.gra	243		110		61		42		20	
	Saccharo.	C.alb	58		69		4		3		0	
		S.cer	45	47	67	70	12	9	3	3	0	0
		C.gla	38		73		12		3		0	
	Archiasco.	S.pom	46	46	61	61	5	5	5	5	0	0
		C.neo	75		68		10		9		3	

Abbreviations: Eurotio. (Eurotiomycetes), Sordario. (Sordariomycetes), Saccharo. (Saccharomycetes), Archiasco. (Archiascomycetes), Basidio. (Basidiomycetes).

Table 3. Cellulolytic enzymes encoded in *T.reesei* genome.

The highest and lowest numbers of entries in each type are indicated in red and blue. *T. reesei* appears in bold.

Cellulase type[a]	CBH1 (Cel7A)	CBH2 (Cel6)	EG1 (Cel7B)	EG2 (Cel5)	EG3 (Cel12)	EG4 (Cel61)	EG5 (Cel45)	
Species[b]								Sum
A.nid	2	2	1	2	1	9	1	18
A.fum	2	1	2	3	3	7	1	19
A.ory	2	1	1	2	2	8	0	16
M.gris	3	2	2	2	3	17	1	30
N.cra	2	2	3	1	0	14	1	23
T.ree	**1**	**1**	**1**	**2**	**1**	**3**	**1**	**10**
F.gra	1	0	1	2	2	13	1	20
C.alb	0	0	0	0	0	0	0	0
S.cer	0	0	0	0	0	0	0	0
C.gla	0	0	0	0	0	0	0	0
S.pom	0	0	0	0	0	0	0	0
C.neo	0	0	0	0	0	1	0	1
P.chr	7	1	2	2	1	14	0	27

[a]Enzyme abbreviations: CBH1 (exocellobiohydrolase I, GH7), CBH2 (exocellobiohydrolase II, GH6), EG1 (endoglucanase I, GH7), EG2 (endoglucanase II, GH5_5), EG3 (endoglucanase III, GH12_1), Cel61 (glycoside hydrolase family GH61).

[b]Species abbreviations: A.nid (*Aspergillus nidulans*), A.fum (*Aspergillus fumigatus*), A.ory (*Aspergillus oryzae*), M.gris (*Magnaporthe grisea*), N.cra (*Neurospora crassa*), T.ree (*Trichoderma reesei*), F.gram (*Fusarium graminearum*), C.alb (*Candida albicans*), S.cer (*Saccharomyces cerevisiae*), C.gla (*Candida glabrata*), S.pom (*Schizosaccharomyces pombe*), C.neo (*Cryptococcus neoformans*), P.chr (*Phanerochaete chrysosporium*).

Table 4. Hemicellulose-degrading enzymes encoded in *T. reesei* genome, arranged by GH family.

Family[a] Species[b]	GH43	GH10	GH11	GH51	GH74	GH62	GH53	GH54	GH67	GH29	GH26	GH95	Total
A.nid	15	3	2	2	2	2	1	1	1	0	3	3	35
A.fum	18	4	3	2	2	2	1	1	1	0	0	2	36
A.ory	20	4	4	3	0	2	1	1	1	0	1	3	40
M.gris	19	5	5	3	1	3	1	1	1	4	0	1	44
N.cra	7	4	2	1	1	0	1	1	1	0	1	0	19
T.ree	**2**	**1**	**4**	**0**	**1**	**1**	**0**	**2**	**1**	**0**	**0**	**4**	**16**
F.gra	16	5	3	2	1	1	1	1	1	1	0	2	34
C.alb	0	0	0	0	0	0	0	0	0	0	0	0	0
S.cer	0	0	0	0	0	0	0	0	0	0	0	0	0
C.gla	0	0	0	0	0	0	0	0	0	0	0	0	0
S.pom	0	0	0	0	0	0	0	0	0	0	0	0	0
C.neo	0	0	0	1	0	0	0	0	0	0	0	0	1
P.chr	4	6	1	2	4	0	1	0	0	0	0	1	19

Note: The highest and lowest numbers of entries in each category are indicated in red and blue. *T. reesei* appears in bold.

[a]Enzymes abbreviated based on CAZyme classification[20].

[b]Species abbreviations: A.nid (*Aspergillus nidulans*), A.fum (*Aspergillus fumigatus*), A.ory (*Aspergillus oryzae*), M.gris (*Magnaporthe grisea*), N.cra (*Neurospora crassa*), T.ree (*Trichoderma reesei*), F.gram (*Fusarium graminearum*), C.alb (*Candida albicans*), S.cer (*Saccharomyces cerevisiae*), C.gla (*Candida glabrata*), S.pom (*Schizosaccharomyces pombe*), C.neo (*Cryptococcus neoformans*), P.chr (*Phanerochaete chrysosporium*).
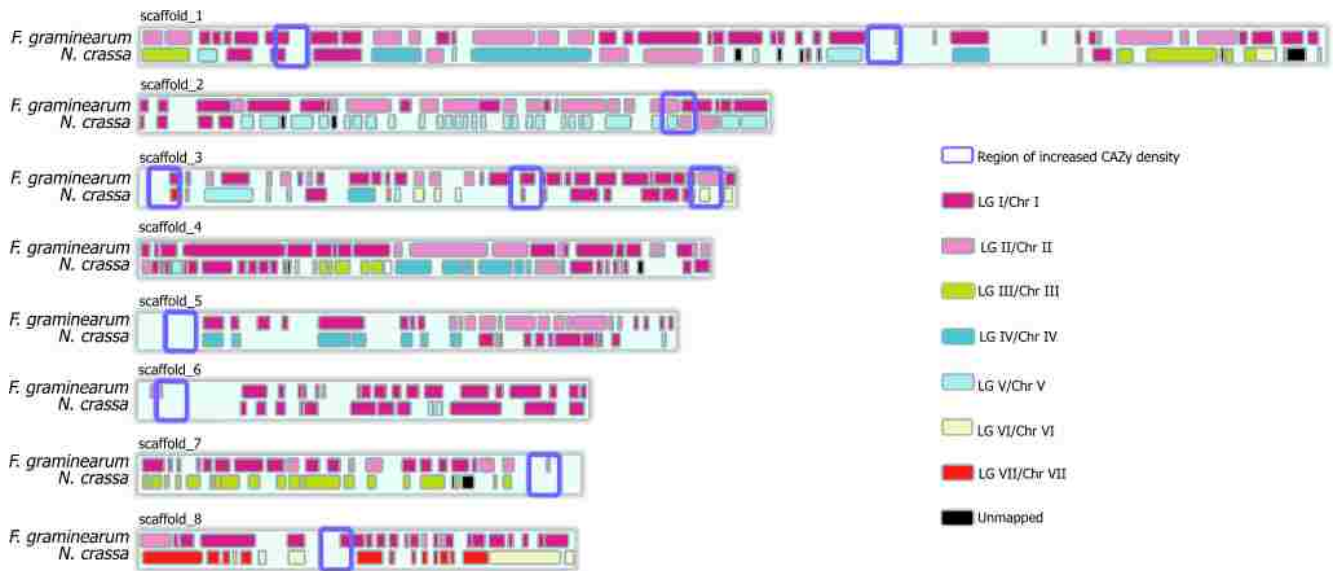
Figure 1. Syntenic blocks mapped to the *Trichoderma reesei* genome from *Fusarium graminearum* and *Neurospora crassa*.

The eight scaffolds displayed comprise approximately one-half the *T. reesei* genome. Small blocks internal to the *T. reesei* genome represent sections of the *T. reesei* genome that share synteny with the *F. graminearum* and *N. crassa* genomes. The calculation of syntenic blocks is described in **Methods**. Syntenic block coloring is by chromosome and linkage group for *F. graminearum* and *N. crassa*, respectively. Blue boxes represent regions of the *T. reesei* genome that have an increased density of genes encoding carbohydrate active enzymes (CAZymes) as described in the text. Overall syntenic comparisons and detailed descriptions of CAZyme blocks are presented in **Supplementary Tables 2** and **7**, respectively.
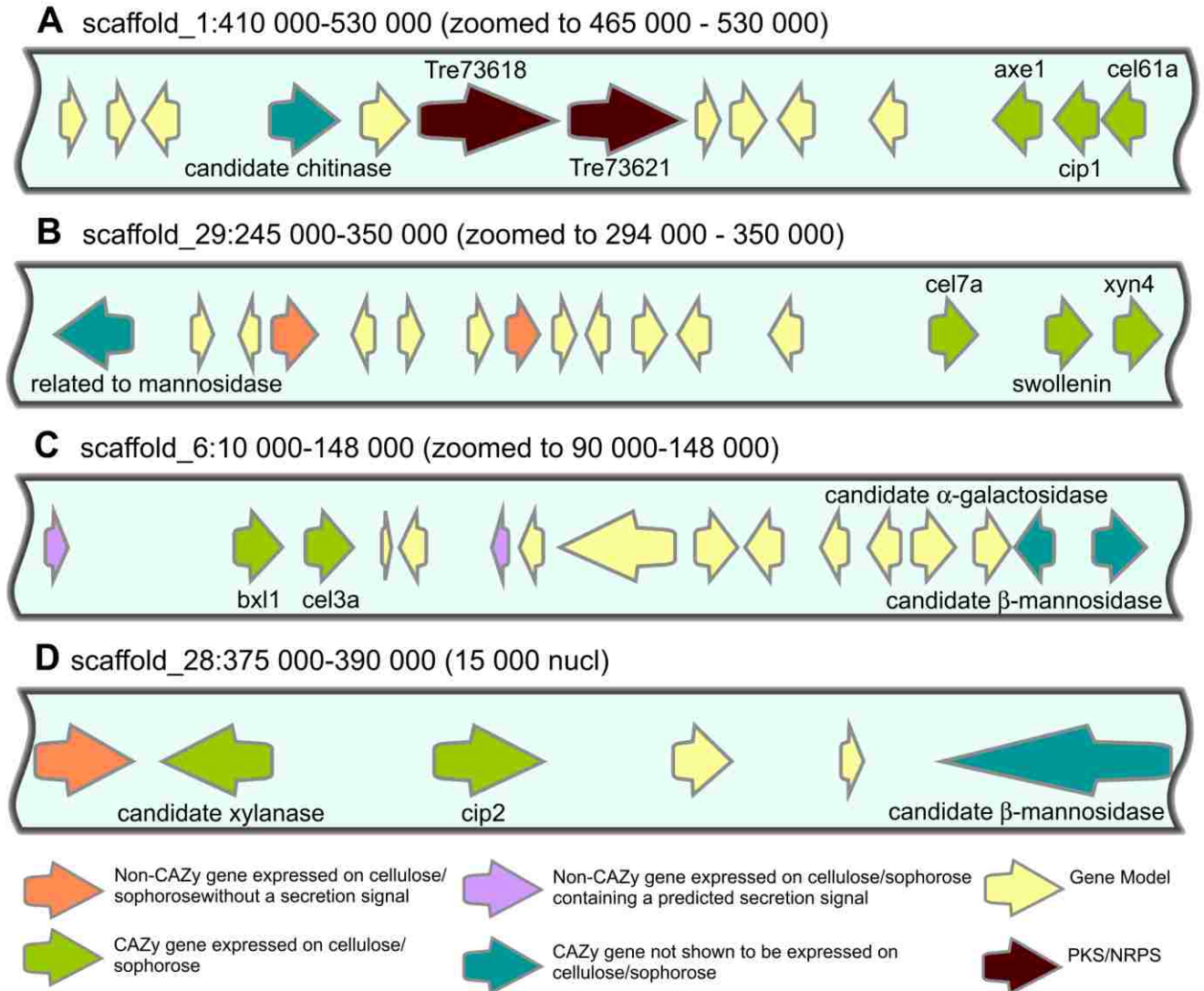
Figure 2. Regions of increased CAZyme density.

Regions of the *T. reesei* genome that contain genes that are co-induced (green arrows) during growth on cellulose and sophorose[42,43]. In all sections of the genome shown in this figure the CAZyme genes are in gaps between syntenic blocks. (A) Region of scaffold _1 containing a candidate chitinase and acetyl xylan esterase (*axe1*), as well as *cip1* (CBM-containing protein identified in Foreman *et al.*[42]) and *cel61a* (endoglucanse IV). The polyketide synthase (PKS) genes in this region are Tr73621 and Tr73618. Tr73621 appears

in a clade with the *lovF* (Lovostatin producing) gene of *Aspergillus terreus*, Tr73618 as described in the text. (B) Region of scaffold_29 showing that cel7a(CBHI), swollenin and xylanase 4 are co-induced.   The upstream orange arrows show co-induction as well on cellulose or sophorose, but are unknown conserved hypotheticals. (C) This region contains *bxl1* (beta-xylosidase) and *cel3a* (β-glucosidase 1) that are co-induced.   Region also contains two unknown genes (purple arrows) containing   predicted secretion signals that are also co-expressed along with *bxl1* and *cel3a*. (D) A candidate xylanase and the CBM containing protein *cip2* (identified in Foreman *et al.*[42]) are adjacent to an unknown gene that is also induced on cellulose/sophorose.

CHAPTER 2

GENOME, TRANSCRIPTOME, AND SECRETOME ANALYSIS OF WOOD DECAY

FUNGUS *POSTIA PLACENTA* SUPPORTS UNIQUE MECHANISMS OF

LIGNOCELLULOSE CONVERSION

Diego Martinez[a,b], Jean Challacombe[a], Ingo Morgenstern[c], David Hibbett[c], Monika Schmoll[d], Christian P. Kubicek[d], Patricia Ferreira[e], Francisco J. Ruiz-Duenas[e], Angel T. Martinez[e], Phil Kersten[f], Kenneth E. Hammel[f], Amber Vanden Wymelenberg[g], Jill Gaskell[f], Erika Lindquist[h], Grzegorz Sabat[i], Sandra Splinter BonDurant[i], Luis F. Larrondo[j], Paulo Canessa[j], Rafael Vicuna[j], Jagjit Yadav[k], Harshavardhan Doddapaneni[k], Venkataramanan Subramanian[k], Antonio G. Pisabarro[l], José L. Lavín[l], José A. Oguiza[l], Emma Master[m], Bernard Henrissat[n], Pedro M. Coutinho[n], Paul Harris[o], Jon Karl Magnuson[p], Scott Baker[p], Kenneth Bruno[p], William Kenealy[q], Patrik J. Hoegger[r], Ursula Kües[r], Preethi Ramaiya[o], Susan Lucas[h], Asaf Salamov[h], Harris Shapiro[h], Hank Tu[h], Christine L. Chee[b], Monica Misra[a], Gary Xie[a], Sarah Teter[o], Debbie Yaver[o], Tim James[s], Martin Mokrejs[t], Martin Pospisek[t], Igor Grigoriev[h], Thomas Brettin[a], Dan Rokhsar[h], Randy Berka[o] and Dan Cullen[f,u]

[a]Los Alamos National Laboratory/Joint Genome Institute, PO Box 1663, Los Alamos, New Mexico 87545; [b]University of New Mexico, Albuquerque, NM 87131; [c]Biology Department, Clark University, Worcester, MA 01610; [d]Research Area Gene Technology and Applied Biochemistry, Institute of Chemical Engineering, Technische Universitat Wien, Getreidemarkt 9/166, A-1060 Vienna, Austria; [e]CIB, CSIC, Ramiro de Maeztu 9, E-28040, Madrid, Spain; [f]Forest Products Laboratory, Madison, WI 53726; [g]Department of Bacteriology, University of Wisconsin, Madison, WI 53706; [h]DOE's Joint Genome Institute, 2800 Mitchell Avenue, Walnut Creek, California 94598; [i]University of Wisconsin Biotechnology Center, Madison, WI 53706; [j]Departamento de Genética Molecular y Microbiología, Facultad de Ciencias Biológicas, Pontifica Universidad Católica de Chile and Millennium Institute for Fundamental and Applied Biology, Santiago, Chile; [k]Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio 45267; [l]Genetics and Microbiology Research Group, Public University of Navarre, 31006 Pamplona, Spain; [m]Chemical Engineering, University of Toronto, Toronto, Ontario, Canada; [n]Architecture et Fonction des Macromolécules Biologiques, UMR 6098, CNRS, Universités d'Aix-Marseille I & II, Case 932, 163 Avenue de Luminy, 13288 Marseille, France; [o]Novozymes Inc, 1445 Drew Avenue, Davis, CA 95618; [p]Pacific Northwest National Laboratory, P.O. Box 999, Richland, WA 99352; [q]Mascoma Inc, Lebanon, NH 03766; [r]Molecular Wood Biotechnology and Technical Mycology, Büsgen-Institute, Georg-August-University Göttingen, Germany; [s]Department of Biology, McMaster University, Hamilton, Ontario, Canada and [t]Department of Genetics and Microbiology, Charles University, Prague, Czech Republic.

Author contributions: D.M., J.C., I.G., E.L., S.L., A.S., H.S., H.T., C.L.C., M.M., G.X., T.B., and D.R. produced genome assembly and automated annotations; D.M., D.H, I.M., C.P.K., E.M., B.H., P.M.C., P.H., S.T., D.Y., and R.B. analyzed carbohydrate active enzymes; D.H., I.M., P.F., F.J.R., A.T.M., P.K., K.E.H., L.L., P.C., R.V., P.H., U.K. and D.C. analyzed oxidoreductases; M.S. and U.K. analyzed signal transduction pathways and mating type; J.Y., H.D. and VS analyzed cytochrome P450s; A.G.P., J.L.L. and J.A.O. analyzed oxidative phosphorylation; J.K.M, W.K. and D.C. analyzed glyoxylate pathway; L.L., and P.C, analyzed iron homeostasis and stress response; S.B., K.B., P.R., T.J., M.M., and M.P.

analyzed other gene sets; A.V.W., J.G., G.S., and D.C. acquired and analyzed mass spectrometry data; D.M., A.V.W., J.G., E.L., S.S.B. and D.C. acquired and analyzed gene expression data; D.M., K.E.H., P.K., R.B. and D.C. integrated analyses and wrote the paper.

To whom correspondence should be addressed. Email: dcullen@wisc.edu, FAX: 608-231-9468.

Classification: Biological Sciences, Microbiology

Manuscript Information: Six pages, three figures

Database deposition: The annotated genome is available on an interactive web portal at http://www.jgi.doe.gov/whiterot. Genome and EST sequence have been deposited and assigned GenBank Accession numbers ABWF00000000 and FL595400-FL633513, respectively.

Brown-rot fungi such as *Postia placenta* are common inhabitants of forest ecosystems and are also largely responsible for the destructive decay of wooden structures. Rapid depolymerization of cellulose is a distinguishing feature of brown-rot, but the biochemical mechanisms and underlying genetics are poorly understood. Systematic examination of the *P. placenta* genome, transcriptome and secretome revealed unique extracellular enzyme systems, including an unusual repertoire of extracellular glycoside hydrolases. Genes encoding exocellobiohydrolases and cellulose-binding domains, typical of cellulolytic microbes, are absent in this efficient cellulose-degrading fungus. When *P. placenta* was grown in medium containing cellulose as sole carbon source, transcripts corresponding to many hemicellulases and to a single putative b-1-4 endoglucanase were expressed at high levels relative to glucose grown cultures. These transcript profiles were confirmed by direct identification of peptides by liquid chromatography-tandem mass spectrometry (LC-MS/MS). Also upregulated during growth on cellulose medium were putative iron reductases, quinone reductase, and structurally divergent oxidases potentially involved in extracellular generation of Fe(II) and H_2O_2. These observations are consistent with a biodegradative role for Fenton chemistry in which Fe(II) and H_2O_2 react to form hydroxyl radicals, highly reactive oxidants capable of depolymerizing cellulose. The *P. placenta* genome resources provide unparalleled opportunities for investigating such unusual mechanisms of cellulose conversion. More broadly, the genome offers insight into the diversification of lignocellulose degrading mechanisms in fungi. Comparisons to the closely related white-rot fungus *Phanerochaete chrysosporium* support an evolutionary shift from white-rot to brown-rot during which the capacity for efficient depolymerization of lignin was lost.

Lignocellulose in vascular plant cell walls is one of the largest sinks for fixed global carbon and is increasingly eyed as a potential feedstock in biofuels and new biomaterials portfolios

48

(1). Relatively few organisms can efficiently convert the recalcitrant polymer blend in lignocellulose to monomeric components (2). The principal exceptions are basidiomycetes, which attack wood through two main decay types called white-rot and brown-rot. Wood-decaying basidiomycetes are essential contributors to carbon cycling in forest soils, and brown-rot fungi are additionally important because they are a major cause of failure in wooden structures. White-rot fungi degrade all components of plant cell walls, including cellulose, hemicellulose and lignin. Although they cannot grow on lignin alone, they have the unique ability to degrade a large proportion of it completely to CO_2 and H_2O. This biodegradative strategy exposes the structural polysaccharides of plant cell walls, thus making them susceptible to hydrolysis by cellulases and hemicellulases. Brown-rot fungi employ a different approach; although they modify lignin extensively, the products remain *in situ* as a polymeric residue (3, 4). Given the incomplete ligninolysis that occurs during brown-rot, it remains unclear how these fungi gain access to plant cell wall polysaccharides. However, it seems probable that the two decay types share at least some mechanisms, because molecular phylogeny, morphological considerations, and substrate preference suggest that brown-rot fungi have repeatedly evolved from white-rot fungi (5). Indeed, the two major experimental organisms for studies of brown-rot, *Postia placenta* and *Gloeophyllum trabeum*, are distantly related species that represent independent origins of brown-rot (5). Any similarities in their decay mechanisms must represent either general mechanisms of wood decay common to white-rot and brown-rot species, or convergently-evolved brown-rot mechanisms. Moreover, *P. placenta* is closely related to the model white-rot fungus, *Phanerochaete chrysosporium*, so comparisons between these species may provide insight into the mechanistic basis of transitions from white-rot to brown-rot.

White-rot fungi produce complex ligninolytic systems that are thought to depend in part on extracellular oxidative enzymes, especially peroxidases, laccases and other oxidases. It remains an open question whether brown-rot fungi possess any of these ligninolytic components. White-rot fungi also secrete complete, synergistically acting cellulase systems that include both endo- and exo-acting enzymes. These exocellobiohydrolases and endoglucanases often share architectures that include separate catalytic and cellulose-binding domains. In contrast, relatively few cellulases have been described in brown-rot fungi (6). It has been long recognized that cellulose depolymerization appears to occur before the substrate porosity has increased enough to admit cellulases (7), and more recent studies (8) have shown that the amorphous regions within cellulose microfibrils are cleaved by *P. placenta* resulting in rapid depolymerization but little weight loss. One possibility consistent with these observations is that brown-rot fungi attack cellulose with low molecular weight oxidants that act in conjunction with a limited set of relatively small cellulases.

The hydroxyl free radical, generated via Fenton chemistry ($H_2O_2 + Fe^{2+} + H^+ \square H_2O + Fe^{3+} + \square OH$), has long been implicated as one of the small oxidants that contributes to polysaccharide depolymerization during brown-rot. Current models for hydroxyl radical participation have been reviewed (6), and typically involve generation of this highly reactive oxidant at or near the substrate. Key requirements for Fenton systems include mechanisms for extracellular H_2O_2 generation and for reduction of Fe^{3+} to Fe^{2+}, which might be accomplished by extracellular fungal metabolites such as hydroquinones, or by extracellular enzymes such as cellobiose dehydrogenase.

We report here analyses of the *P. placenta* draft genome together with transcript profiles and mass spectrometric identification of extracellular proteins. Consistent with a unique

50

strategy for cellulose degradation, we observed a dramatic absence of conventional cellulase genes and most class II fungal peroxidases, and a rich diversity of genes potentially supporting generation of extracellular reactive oxygen species.

RESULTS

Carbohydrate Active Enzymes

Given the well known efficiency with which brown-rot fungi rapidly depolymerize and degrade cellulose, the *P. placenta* genome revealed remarkably few, if any, conventional cellulases. Of 17,173 proteins predicted in the dikaryotic genome, 242 unique genes encode potential carbohydrate-active enzymes ((9); http://www.cazy.org), of which 228 (94%) have at least one potential ortholog (BLASTP bit score $\geq$100) in *P. chrysosporium*. These putative CAZY genes include 144 glycoside hydrolases (GH), 10 carbohydrate esterases (CE), 75 glycosyltransferases (GT), 7 expansin-like proteins (EXPN), and 6 polysaccharide lyases (PL) (complete CAZY list in SI Table 1 within NCBI GEO accession 12540. In distinct contrast to all cellulolytic fungal aerobes, exocellobiohydrolases CBH2 (GH6) and CBH1 (GH7), as well as cellulose-binding endoglucanases are missing in the *P. placenta* genome (Fig.1). Also absent are family 1 carbohydrate binding modules (CBM1). These highly conserved cellulose-binding domains are fused to functionally diverse CAZYs in a wide range of cellulolytic microbes. Surprisingly then, the repertoire of recognizable cellulolytic enzymes in *P. placenta* appears limited to just two potential endoglucanases (1,4-b-glucanases) and several β-glucosidases. In contrast to cellulolytic saprophytes (e.g. *Trichoderma reesei*, *Aspergillus spp.* or *Neurospora crassa*) and aggressive plant pathogens (e.g. *Fusarium graminearum* or *Magnaporthe grisea*), the overall number and distribution of GHs in *P. placenta* are similar to those in the ectomycorrhizal symbiont *Laccaria bicolor*, the human pathogen *Cryptococcus neoformans* and the biotrophic plant

51

pathogen *Ustilago maydis* (SI Table 2). Phylogenetic analyses of *P. placenta* and *P. chrysosporium* genomes indicate that the transition from white-rot to brown-rot has been associated with multiple independent reductions including the GH families 6, 7, 10, 11 and 61 (Figs. 1 and 2; SI Table 2) Thus, the transition from white-rot to brown-rot has been associated with multiple independent reductions in the GH families.

Microarrays representing 12,438 unique alleles were used to examine *P. placenta* transcript levels in basal salts medium containing either glucose or wood-derived microcrystalline cellulose as the sole carbon sources. In total, 290 gene models showed >2 fold transcript accumulation, and of these, 235 increased in cellulose medium and 35 increased in glucose medium (NCBI GEO accession 12540 SI Table 3). Transcripts of 99 GH-encoding genes significantly increased (P<0.01) in the cellulose medium, and of these, 18 increased >2 fold (Fig. 3). Twenty-one GH transcripts significantly increased in the glucose-containing medium, but none exceeded a two-fold change. In addition, shotgun liquid chromatography coupled tandem mass spectrometry (LC-MS/MS) identified 26 specific CAZYs in the extracellular fluid of *P. placenta* grown in basal salts supplemented with ball-milled aspen wood, microcrystalline cellulose, or cotton (SI Table 4). The CAZY genes expressed in cellulose included laminarases, chitinases, and various hemicellulases (endoxylanases, β-xylosidases, L-α-arabinofuranosidases, endo-β-mannanases and β-mannosidases). It is unclear whether any of these enzymes could directly attack crystalline cellulose.

Extracellular H$_2$O$_2$ Generation

Gene models potentially supporting Fenton chemistry through the generation of extracellular H$_2$O$_2$ include copper radical oxidases and GMC oxidoreductases (SI Table 5). Results summarized here focus on those genes with expression patterns that are consistent with a

role in cellulose depolymerization, and detailed information for all genes is available in SI Table 6 within GEO 12540.

On the basis of overall sequence similarity to *P. chrysosporium* glyoxal oxidase (GLOX) and conservation of catalytic residues (10), three *P. placenta* models were identified as copper radical oxidases (CROs). GLOX is one of 7 CROs in *P. chrysosporium* and physiologically coupled to lignin peroxidase (LiP) via H_2O_2 generation. Of particular relevance to potential Fenton systems, CRO genes encoding proteins Ppl56703 and Ppl130305 are upregulated in microcrystalline cellulose, and Ppl56703 peptides were detected in aspen-grown cultures. The *P. chrysosporium cro1* and *glx1* genes are not closely related, and they do not have orthologs in *P. placenta,* which suggests that there have been two independent losses of these CRO lineages in *Postia* (Fig. 2). Ppl56703 is orthologous to the *cro3-4-5* lineage in *P. chrysosporium*, which therefore represents a *Phanerochaete*-specific expansion of the gene family. As in the case of the GHs, evolution of brown-rot is associated with a reduced diversity of CROs.

Catalytically distinct from CROs, GMC oxidoreductases (InterPro IPR000172) included various alcohol and sugar oxidases. Among the former, *P. placenta* protein model Ppl118723 is similar to *G. trabeum* methanol oxidase (DQ835989) (> 85% amino acid identity over full length). Recent immunolocalization studies strongly implicate the *G. trabeum* alcohol oxidase as a source of H_2O_2 (11) to support Fenton chemistry. Suggesting a similar role in *P. placenta*, microarray analysis revealed high transcript levels and a sharp increase in transcription of the gene encoding Ppl118723 in cellulose-grown cultures relative to non-cellulolytic cultures. Comparatively high transcript levels in cellulose- and glucose-grown cultures were also observed for genes encoding Ppl128830 and Ppl108489, models

tentatively identified as glucose-1-oxidases based on conserved key residues (12). Peptides corresponding to these putative *gox* genes were detected in extracellular filtrates (SI Tables 6 and 11 within GEO 12540). Aryl-alcohol oxidase, an extracellular GMC oxidoreductase cooperating with aryl-alcohol dehydrogenases for continuous peroxide supply in some white-rot fungi (12) does not seem to be involved in cellulose attack by *P. placenta* since the corresponding models were not or only slightly upregulated. Another extracellular GMC oxidoreductase, pyranose-2-oxidase, has been implicated in lignocellulose degradation in *P. chrysosporium* (13), but no orthologs were detected in *P. placenta*.

Iron Reduction and Homeostasis

Protein model Ppl124517 was identified as a putative quinone reductase (QRD). In the brown-rot fungus *G. trabeum*, a QRD may drive extracellular Fenton systems via redox cycling of secreted fungal quinones (6). Transcription of the *P. placenta* QRD gene was significantly upregulated in cellulose medium (GEO 12540 SI Table 3), which is consistent with a role for cellulolytic Fenton chemistry involving quinone redox-cycling. In this connection, upregulation of the genes encoding phenylalanine ammonia lyase (Ppl112824) and a putative quinate transporter (Ppl44553) may also be relevant by virtue of their respective roles in the biosynthesis and transport of essential quinones.

In addition to hydroquinone-based iron reduction systems, low molecular weight glycoproteins (GLPs) that can act as iron reductases have been hypothesized as components of extracellular Fenton systems in *G. trabeum* and *P. chrysosporium* (14). Four *P. placenta* models show significant similarity (>48% amino acid identity) to *P. chrysosporium glp1* and *glp2*, and the gene encoding Ppl128974 is significantly upregulated on microcrystalline cellulose medium (GEO 12540 SI Table 7). Sequence corresponding to

another fungal protein implicated in Fe^{3+} reduction, CDH (6), appears to be absent in *P. placenta*.

In addition to its pivotal role in a wide range of cellular processes, iron homeostasis must play a central role in modulating a functioning Fenton system. The *P. placenta* genome features numerous genes potentially involved in iron transport and redox state (GEO 12540 SI Table 7). In addition to 7 ferric reductases, two iron permeases were identified one of which lies immediately downstream from a canonical yeast ferroxidase ortholog (Fet3). Transcripts of these adjacent genes were among the most highly upregulated in cellulose medium (GEO 12540 SI Table 3).

Modification of Lignin and Other Aromatic Compounds

Genes encoding the class II secretory peroxidases lignin peroxidase (LiP), manganese peroxidase (MnP) and versatile peroxidase (VP) were not detected in the *P. placenta* genome (SI Table 5). Peroxidase model Ppl44056 lacks residues involved in Mn^{2+} binding and oxidation of aromatic compounds (15), and superimposition of protein models strongly suggests that Ppl44056 is a low redox potential peroxidase (SI Fig. 1). Consistent with this structural evidence, phylogenetic analyses of class II peroxidase genes from *Postia, Phanerochaete*, and other fungal genomes suggest that Ppl44056 is not closely related to LiP and MnP, but is part of an assemblage of "basal peroxidases" that includes the novel peroxidase (NoP) of *P. chrysosporium,* and peroxidases from *Coprinopsis cinerea* and *L. bicolor* (Fig. 2) (16). The backbone of the class II peroxidase phylogeny is not strongly supported, but analyses of broadly sampled datasets (16), suggest that the LiP and MnP gene lineages of *P. chrysosporium* were independently derived from the basal peroxidases

prior to the divergence of *Postia* and *Phanerochaete*. If so, then the absence of LiP and MnP in *P. placenta* may reflect additional instances of gene loss.

Laccases have been suggested to play a role in lignin modification by white-rot fungi, but have not previously been demonstrated in brown-rot fungi. The precise role of these enzymes remains uncertain, but numerous studies have demonstrated laccase-catalyzed oxidation of phenolic and nonphenolic lignin model substrates particularly in the presence of low molecular weight mediators. The results from *P. placenta* belie the usual picture of brown-rot in that models Ppl62097 and Ppl111314 are likely laccases *sensu stricto* (17) (Fig.2). Transcript levels of the genes encoding Ppl89382 and Ppl111314 appear differentially regulated by decreasing slightly (-1.08-fold) and increasing (+2.29-fold), respectively, on cellulose medium relative to glucose medium (GEO 12540 SI Table 7). These enzymes could contribute to hydroxyl radical generation by oxidizing hydroquinones as described (18). Interestingly, laccase genes are absent from the genome of *P. chrysosporium* (19), suggesting that laccase (*sensu stricto*) is not a core component of fungal wood decay mechanisms, and is certainly not essential for white-rot.

Other upregulated genes potentially involved in quinone redox-cycling, and oxidation of lignin derived products include those encoding "polyphenol oxidase" (Ppl114245), i.e. tyrosinase or catechol oxidase related to typical laccases, and various oxidoreductases of uncertain function (Ppl107061, Ppl28683, Ppl34850, Ppl61437, Ppl24981) (SI Table 3).

Oxalate Metabolism

In addition to pH effects on a wide range of enzymes, extracellular accumulation of oxalate by *P. placenta* may affect ferric iron availability and thereby impact hydroxyl radical

formation (20) reviewed in (6). A metabolic shunt between the citric acid and glyoxylate cycles is central to oxalic acid accumulation by the brown-rot fungus *Fomitopsis palustris* (21). Analysis of the *P. placenta* genome demonstrates a functional glyoxylate shunt and substantially extends our understanding of the number, structure, and transcription of key genes (SI Fig.2; SI text; GEO 12540 SI Table 8).

Cytochrome P450 Monooxygenases

P450s have various roles in secondary metabolism and thought to be involved in biodegradation of lignin as well as various xenobiotic compounds. The *P. placenta* genome features an impressive set of 236 P450 genes (SI text, GEO 12540 SI Fig. 3), compared to 149 in *P. chrysosporium*, and expansions of certain families (CYP64, CYP503, CYP5031 and CYP617) were observed. Genes encoding Ppl110015 (CYP53) and Ppl128850 (CYP503) were significantly upregulated in cellulose medium (GEO 12540 SI Table 3). The former is highly conserved in fungi and thought to catalyze benzoate hydroxylation.

Other

The genome was systematically examined for genes involved in oxidative phosphorylation, stress-related genes, signal transduction and regulatory genes, particularly those potentially controlling glycoside hydrolase expression and mating type (complete listings and analysis in SI text, GEO 12540 Figs 5-7).

DISCUSSION

Analysis of the *P. placenta* genome elucidated a repertoire of genes and expression patterns distinct from those of other known cellulose-degrading microbes. The overall number of CAZY-encoding genes in *P. placenta*, 242, is not particularly low, and among

these, the number of glycosyl transferases, 75, is fairly typical. However, the genome completely lacks cellulose-binding domains and the number of GHs is relatively low owing in part to the paucity of cellulases. No exocellobiohydrolases and only two potential b-1,4 endoglucanase genes were identified. One putative EG (Ppl115648) is expressed at relatively high levels.

Comparisons with genomes of other cellulolytic microbes reveal a strikingly distinct set of glycoside hydrolases in the *P. placenta* genome. Among aerobes, only the cellulolytic gliding bacterium, *Cytophaga hutchinsonii* lacks exocellobiohydrolases and endoglucanases fused to cellulose-binding domains (22). The precise mechanism employed by *C. hutchinsonii* is somewhat mysterious, but it has been suggested that cellulose chains are peeled away from the polymer and transported into the periplasm (23). There, non-processive endoglucanases might readily degrade the cellulose. Such a mechanism seems unlikely in *P. placenta* because all evidence suggests that cellulose depolymerization by brown-rot fungi occurs at a distance from the advancing hyphae. In contrast, *C. hutchinsonii* is in direct contact with cellulose.

Possibly, the CBM-less b-1-4-endoglucanase Ppl115648, which is clearly expressed in cellulose-containing media (Fig. 3), may possess processive activity that enables it to liberate the cellobiose that β-glucosidases then hydrolyze to assimilable glucose. Indeed, the accumulation of putative b-glucosidase transcripts and the corresponding proteins that we observed are consistent with the availability of cellobiose in our cultures. Precedents for crystalline cellulose hydrolysis by b-1,4-endoglucanases within GH family 5 have been reported (24, 25), but it seems unlikely that the Ppl115648 endoglucanase alone can account for the efficient cellulose depolymerization by *P. placenta*. Other GHs and/or

hypothetical proteins, perhaps some of those expressed in microcrystalline cellulose cultures (Fig. 3; GEO 12540 SI Table 1), may be necessary for the complete breakdown of cellulose. Heterologous expression of *P. placenta* GH-encoding genes followed by biochemical characterization of the purified proteins may resolve this question.

Many investigations of white-rot and brown-rot mechanisms have implicated the participation of low molecular weight oxidants, particularly Fenton-generated hydroxyl radicals. As recently reviewed (6), three somewhat overlapping mechanisms of oxidative degradation have been advanced. One view emphasizes the importance of CDH. In the case of *P. placenta*, CDH is absent. Another view invokes the role of low molecular weight glycopeptides that catalyze extracellular iron reduction. Initially identified in *P. chrysosporium* (14), potential orthologs of these glycopeptide-encoding genes were identified in *P. placenta*, and in one case, increased transcript levels were observed in cellulose medium. Accordingly, a role for these glycoproteins in a *P. placenta* Fenton system is possible. The third mechanism involves extracellular quinone redox cycling (26). Evidence supporting this system includes cellulose induction of genes encoding QRD, quinate transporter, phenylalanine ammonia lyase and laccase. However, the importance of hydroquinone-driven Fenton chemistry in *P. placenta* remains unclear because this fungus secretes high levels of oxalate (27), and Fe^{3+}-oxalate chelates are poorly reducible by hydroquinones (28).

The elevated expression in cellulose medium of Fet3 and Ftr1, components of the high affinity iron uptake system, may be at least partially explained by such chelates. While cellulose itself may sequester Fe^{3+} (29), the generation of Fe^{3+}–oxalate and potentially other redox active iron-chelates might also contribute to lower the effective concentration of

bioavailable iron that is accessible to the organism. Thus, cellulolytic conditions might turn on the high-affinity iron uptake system to ensure proper levels of intracellular iron.

Also compatible with Fenton mechanisms is the observed cellulose-induced expression of structurally divergent oxidases (e.g. copper radical oxidases, glucose-1-oxidases and methanol oxidases) and putative iron reductases. Given the significant number of secreted hemicellulases observed, wood decay by *P. placenta* likely involves attack by oxidative and hydrolytic mechanisms. Elevated hemicellulase expression may reflect increased substrate exposure and availability, relative to cellulose and lignin, especially early in the decay process. Products of the hydrolytic attack could in turn serve as candidate substrates for copper radical oxidases and GMC oxidoreductases, thereby generating extracellular H_2O_2. Similarly, methanol released via demethoxylation of lignin (3, 4) may play an important role in H_2O_2 generation as a substrate for methanol oxidase. Such a role is consistent with our observed expression patterns and with previous investigations with *G. trabeum* (11). Of course hydroxyl radical may also play an important role early in decay, and it has been demonstrated to preferentially attack hemicellulose in wood (30). Interestingly, •OH attack on cellulose oxidizes chain ends (31) and the depolymerized material becomes less amenable to cellulase action (30), providing a plausible explanation for the lack of exocellobiohydrolase genes in this fungus.

Comparison of the *P. placenta* and *P. chrysosporium* genomes indicates that the derivation of brown-rot is characterized largely by the contraction or loss of multiple gene families that are thought to be important in typical white-rot, such as cellulases, LiPs, MnPs, CROs, CDH, and pyranose-2-oxidase. This general pattern of simplification is consistent with the view (32) that brown-rot fungi, having evolved novel mechanisms for initiating cellulose

60

depolymerization, have cast off much of the energetically costly lignocellulose-degrading apparatus that is retained in white-rot fungi, such as *P. chrysosporium*.

MATERIALS AND METHODS

Genome Sequencing, Assembly and Annotation

A pure whole genome shotgun approach was used to sequence *P. placenta* strain MAD-698-R (USDA, Forest Mycology Center, Madison, WI). The 7.2X coverage assembly was produced from sequenced paired reads using JAZZ assembler. Using an array of gene predictors in the JGI annotation pipeline, a total of 17,173 gene models were predicted and annotated for this dikaryotic fungus. Predicted genes, supporting evidence, annotations, and analyses are available through interactive visualization and analysis tools from the JGI genome portal (http://genome.jgi-psf.org/Pospl1/Pospl1.home.html). Detail regarding the assembly, repetitive elements, ESTs and annotation, are provided separately (SI text).

Mass Spectrometry

Soluble extracellular protein was concentrated from shake cultures containing ball-milled aspen, microcrystalline cellulose (Avicel) or de-waxed cotton as previously reported (33). Sample preparation and LC-MS/MS analysis were performed as described (www.biotech.wisc.edu/ServicesResearch/MassSpec/ingel.htm.). Peptides were identified using a Mascot search engine (Matrix Science, London, UK) against protein sequences of 17,173 predicted gene models described above. Complete listings of CAZYs and oxidative enzymes, including peptide sequences and scores, are provided in SI text and in NCBI's GEO under series accession GSE12540 SI Table 11).

Expression Microarrays

Roche NimbleGen (Madison, WI) arrays were designed to assess expression of 12,438 genes during growth on microcrystalline cellulose or on glucose as sole carbon sources. The corresponding set of coding regions was manually annotated to include only the 'best allelic model' among CAZY-encoding genes (GEO GSE12540 SI Table 1). Methods are detailed in SI text, and all data deposited under GEO accession GSE12540.

Acknowledgements

Supplementary information (SI) is available online at
http://www.pnas.org/content/106/6/1954/suppl/DCSupplemental.

References

1.	DOE US (2006) *Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda, DOE/SC-0095*, (DOE).

2.	Eriksson K-EL, Blanchette RA, & Ander P (1990) *Microbial and enzymatic degradation of wood and wood components* (Springer-Verlag, Berlin).

3.	Niemenmaa O, Uusi-Rauva A, & Hatakka A (2007) Demethoxylation of [O(14)CH (3)]-labelled lignin model compounds by the brown-rot fungi *Gloeophyllum trabeum* and *Poria (Postia) placenta*. *Biodegradation* 19:555-565.

4.	Yelle DJ, Ralph J, Lu F, & Hammel KE (2008) Evidence for cleavage of lignin by a brown rot basidiomycete. *Environ Microbiol* 10:1844-1849.

5.	Hibbett DS & Donoghue MJ (2001) Analysis of character correlations among wood decay mechanisms, mating systems, and substrate ranges in homobasidiomycetes. *Syst. Biol.* 50(2):215-242.

6.	Baldrian P & Valaskova V (2008) Degradation of cellulose by basidiomycetous fungi. *FEMS Microbiol Rev* 32(3):501-521.

7.	Cowling EB & Brown W (1969) Structural features of cellulosic materials in relation to enzymatic hydrolysis. *Cellulases and Their Applications*, eds Hajny GJ & Reese ET (American Chemical Society Advances in Chemistry Series 95, Washington, DC), pp 152-187.

8.	Kleman-Leyer K & Kirk TK (1992) Changes in the molecular size distribution of cellulose during attack by white-rot and brown-rot fungi. *Appl. Environ. Microbiol.* 58:1266-1270.

9.	Henrissat B (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J* 280 (Pt 2):309-316.

10.	Whittaker MM, Kersten PJ, Cullen D, & Whittaker JW (1999) Identification of catalytic residues in glyoxal oxidase by targeted mutagenesis. *J Biol Chem* 274(51):36226-36232.

11.	Daniel G*, et al.* (2007) Characteristics of *Gloeophyllum trabeum* alcohol oxidase, an extracellular source of H_2O_2 in brown rot decay of wood. *Appl Environ Microbiol* 73(19):6241-6253.

12.	Varela E, Martinez JM, & Martinez AT (2000) Aryl-alcohol oxidase protein sequence: a comparison with glucose oxidase and other FAD oxidoreductases. *Biochim Biophys Acta* 1481(1):202-208.

13.	de Koker TH, Mozuch MD, Cullen D, Gaskell J, & Kersten PJ (2004) Pyranose 2-oxidase from *Phanerochaete chrysosporium*: isolation from solid substrate, protein purification, and characterization of gene structure and regulation. *Appl. Environ Microbiol* 70:5794-5800.

14.   Tanaka H*, et al.* (2007) Characterization of a hydroxyl-radical-producing glycoprotein and its presumptive genes from the white-rot basidiomycete *Phanerochaete chrysosporium. J Biotechnol* 128(3):500-511.

15.   Martinez AT (2002) Molecular biology and structure-function of lignin-degrading heme peroxidases. *Enzyme Microb Technol* 30:425-444.

16.   Morgenstern I, Klopman S, & Hibbett DS (2008) Molecular evolution and diversity of lignin degrading heme peroxidases in the Agaricomycetes. *J Mol Evol* 66(3):243-257.

17.   Hoegger PJ, Kilaru S, James TY, Thacker JR, & Kües U (2006) Phylogenetic comparison and classification of laccase and related multicopper oxidase protein sequences. *FEBS J* 273(10):2308-2326.

18.   Guillen F, Gomez-Toribio V, Martinez MJ, & Martinez AT (2000) Production of hydroxyl radical by the synergistic action of fungal laccase and aryl alcohol oxidase. *Arch Biochem Biophys* 383(1):142-147.

19.   Martinez D*, et al.* (2004) Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nature Biotechnol* 22:695-700.

20.   Varela E & Tien M (2003) Effect of pH and oxalate on hydroquinone-derived hydroxyl radical formation during brown rot wood degradation. *Appl Environ Microbiol* 69(10):6025-6031.

21.   Munir E, Yoon JJ, Tokimatsu T, Hattori T, & Shimada M (2001) A physiological role for oxalic acid biosynthesis in the wood-rotting basidiomycete *Fomitopsis palustris. Proc Natl Acad Sci USA* 98(20):11126-11130.

22.   Xie G*, et al.* (2007) Genome sequence of the cellulolytic gliding bacterium *Cytophaga hutchinsonii. Appl Environ Microbiol* 73(11):3536-3546.

23.   Wilson DB (2008) Three microbial strategies for plant cell wall degradation. *Ann N Y Acad Sci* 1125:289-297.

24.   McCarter SL*, et al.* (2002) Exploration of cellulose surface-binding properties of *Acidothermus cellulolyticus* Cel5A by site-specific mutagenesis. *Appl Biochem Biotechnol* 98-100:273-287.

25.   Tsai CF, Qiu X, & Liu JH (2003) A comparative analysis of two cDNA clones of the cellulase gene family from anaerobic fungus *Piromyces rhizinflata. Anaerobe* 9(3):131-140.

26.   Cohen R, Suzuki MR, & Hammel KE (2004) Differential stress-induced regulation of two quinone reductases in the brown rot basidiomycete *Gloeophyllum trabeum. Appl Environ Microbiol* 70(1):324-331.

27.   Kaneko S, Yoshitake K, Itakura S, Tanaka H, & Enoki A (2005) Relationship between production of hydroxyl radicals and degradation of wood, crystalline cellulose, and

lignin-related compound or accumulation of oxalic acid in cultures of brown-rot fungi. *J Wood Sci* 51:262-269.

28. Jensen KA, Jr., Houtman CJ, Ryan ZC, & Hammel KE (2001) Pathways for extracellular Fenton chemistry in the brown rot basidiomycete *Gloeophyllum trabeum. Appl Environ Microbiol* 67(6):2705-2711.

29. Xu G & Goodell B (2001) Mechanisms of wood degradation by brown-rot fungi: chelator-mediated cellulose degradation and binding of iron by cellulose. *J Biotechnol* 87(1):43-57.

30. Ratto M, Ritschkoff A, & Viikari L (1997) The effect of oxidative pretreatment on cellulose degradation by *Poria placenta* and *Trichoderma reesei. Appl Microbiol Biotechnol* 48:53-57.

31. Kirk TK, Ibach R, Mozuch MD, Conner AH, & Highley TL (1991) Characteristics of cotton cellulose depolymerized by a brown-rot fungus, by acid, or by chemical oxidants. *Holzforschung* 45:239-244.

32. Worrall JJ, Anagnost SE, & Zabel RA (1997) Comparison of wood decay among diverse lignicolous fungi. *Mycologia* 89:199-219.

33. Vanden Wymelenberg A*, et al.* (2006) Computational analysis of the *Phanerochaete chrysosporium* v2.0 genome database and mass spectrometry identification of peptides in ligninolytic cultures reveals complex mixtures of secreted proteins. *Fungal Genetics and Biology* 43:343-356.
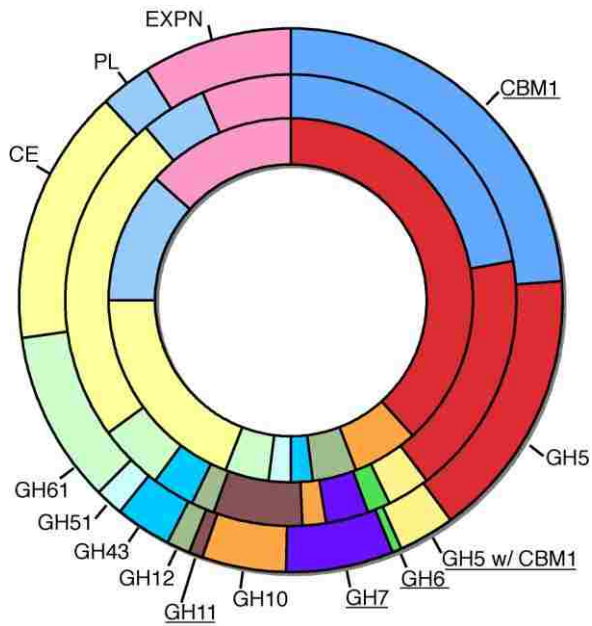
Figure 1. Distribution of various CAZymes in *P. placenta* (inner ring), *T. reesei* (middle ring), and *P. chrysosporium* (outer ring).

Abbreviations: CBM1, family 1 carbohydrate binding modules; GH#, modules within individual glycoside hydrolase families; GH5 (CBM1), glycoside hydrolase family 5 modules associated with family 1 carbohydrate binding modules; GT, glycosyl transferases; CE, carbohydrate esterases; PL, polysaccharide lyases; EXPN, expansin-related proteins. Enzymes not found in *P. placenta* are underlined. Comparisons with additional species are list listed in SI Table1.
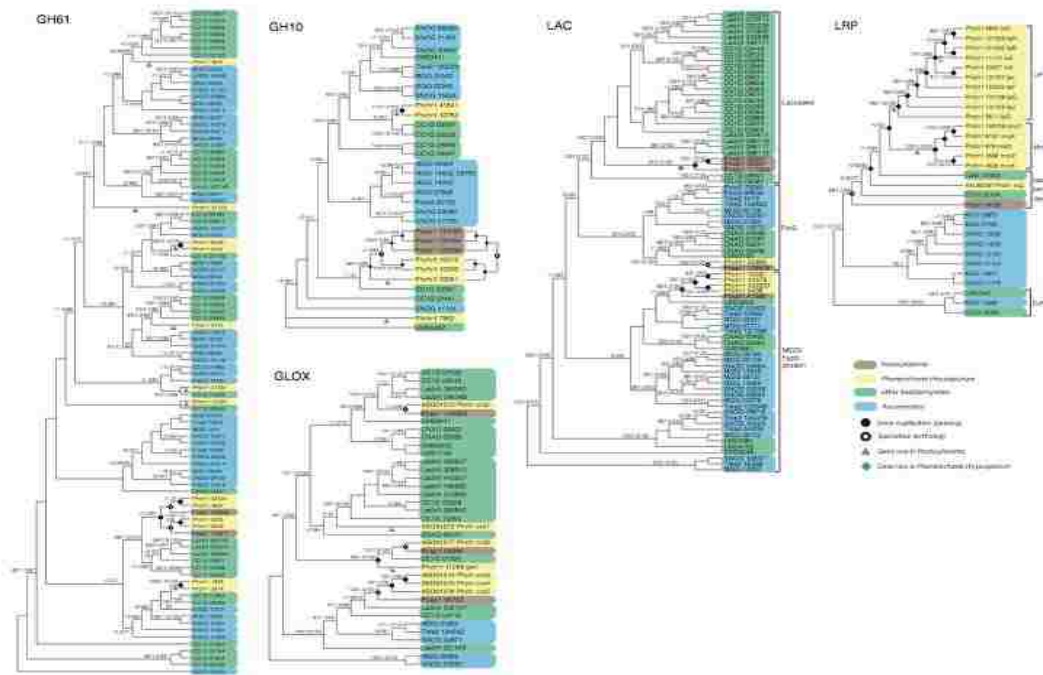
Figure 2. Phylogenies of glycoside hydrolase (GH 61, GH10), glyoxal oxidase/copper radical oxidase (GLOX), laccase (LAC) and related multicopper oxidase, and low redox peroxidase (LRP) and related class II fungal peroxidases, from complete genomes of *P. placenta* (Pospl1), *P. chrysosporium* (Phchr1), *C. cinerea* (CC1G)*, L. bicolor* (Lacbi1)*, C. neoformans* (CNAG)*, U. maydis* (UM), *M. grisea* (MGG)*, Stagonospora nodorum* (SNOG)*, T. reesei* (Trire2) and *Pichia stipitis* (Picst3).

Datasets were assembled using BLAST (with qUniProtKB query sequences Q5XXE5, O60206, P36218, Q00023, O14405, Q01772, Q12718), with a cut-off value of E-06. Parsimony (PAUP* 4.0; 10,000 heuristic searches, 1000 bootstrap replicates), maximum likelihood (RAxML; 1000 bootstrap replicates, with models suggested by ProtTest), and Bayesian (MrBayes v3.1.2; two runs of four chains, 10 million generations each, with mixed protein models) support values are indicated in the order MP/PP/ML. Topologies shown are from Bayesian phylogenetic analyses. An alternative topology from parsimony analysis is shown for part of the GH10 phylogeny. Inferred gene losses, duplication events (paralogy), and speciation events (orthology) are indicated within *Postia* and *Phanerochaete* only.
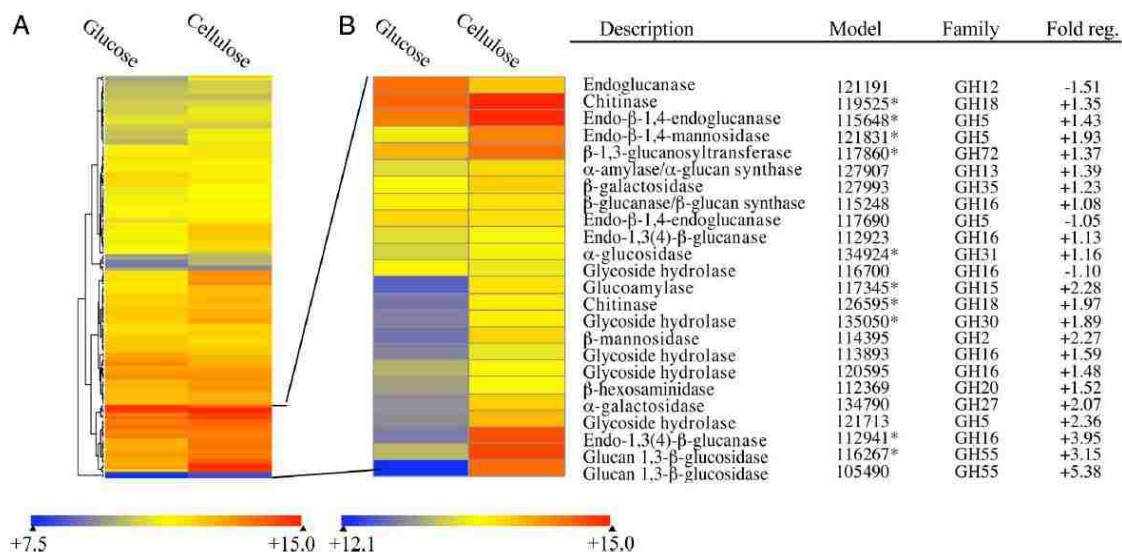
Figure 3. Expression profile of 144 glycoside hydrolase-encoding genes in media containing glucose versus microcrystalline cellulose as sole carbon sources (Part A). In Part B, a cluster of 24 of highly expressed genes is expanded and the color scale re-calibrated to illustrate differences in transcript accumulation.

Expression ratios were derived from comparisons of glucose-grown versus celluose-grown mycelia. Analysis is based on three full biological replicates per culture medium. Quantile normalization and robust multi-array averaging (RMA) were applied to the entire dataset. ANOVA showed 120 GH-encoding genes, including all 24 above, were significantly regulated (P<0.01). Reciprocals of ratios <1.0 are multiplied by -1. Asterisks indicate proteins identified by LC-MS/MS. A detailed listing of all CAZYs with statistical analyses of expression data is presented in SI Table 1.

ADDENDUM 1: ORTHOLOGS AND KA/KS ANALYSIS OF ALLELES

Previous whole genome comparisons have revealed many genes that are unique to particular species and are thought to be important in adaptation[1,2]. By comparing the gene content of *Postia placenta* with that of other basidiomycete genomes, we were able to identify genes that were specific to *P. placenta*. We identified gene products that appeared in all genomes or were unique by aligning the proteins from *P. placenta*, *Laccaria bicolor*[3] and *Phanerochaete chrysosporium*[1] with the BLAST algorithm. We set a liberal cutoff of a bit score of 50 or greater, and the coverage of both subject and query (defined as quotient of the alignment length and the predicted protein length) to a minimum of 20 percent, to be considered a match. We used the condensed version of the *P. placenta* proteome, which contains only one protein from each allele pair present (see Materials and Methods in the main portion of Chapter 2).

The Results of the three-way comparison is shown in **Figure A1.1**. In each section of the Venn diagram that contains proteins from two or more genomes, we show the number of proteins that came from each genome. For example, the center of the Venn diagram contains proteins that were found in all three genomes. However, *Laccaria bicolor* had the largest number of proteins with similarity in the other two genomes. This indicates a contraction or expansion in the number of proteins that are related to the other two basidiomycetes. This result is not surprising, as it reflects the total number of genes in each genome; *L. bicolor* contains almost twice as many genes as the other two fungi. Also, from the diagram we can see that the largest areas of gene expansion involve unique genes as well as those conserved in all three genomes.

We were interested in the functions of genes that are unique to *P. placenta*, as these may be involved in species specific abilities. To determine the function of the unique genes in *P. placenta,* we used runInterproscan [4] to annotate the predicted proteins. As this search can also double as a search for potentially novel protein domain combinations, we have allowed multiple Interpro domain assignments for a single protein to remain as multiple annotations and included the counts in the table, thus preserving potentially unique domain combinations. By comparing the number of Interpro domains from each area of the Venn in **Figure A1.1** for the *P. placenta* genome, we found that the number of predicted proteins with identifiable Interpro domains was far lower in the unique genes category of *P. placenta*, as this group contained only 20.6% of genes products with identifiable Interpro domains. In contrast, 67.0% of the *P. placenta* genes in the group containing similarity to proteins of all three genomes contained Interpro matches. While this difference is striking, it is not surprising; as we discussed above, lineage-specific genes usually contain a large number of unknown genes. Genes that are required for all eukaryotic cellular life encode functions that are well studied in model organisms, which usually dominate databases such as Interpro.

We further analyzed the Interpro results for the possibility that genes with similar function may be differentially represented among the various Venn categories in **Figure A1.1**. From these data, we saw intriguing patterns emerge. We first analyzed the proteins in *P. placenta* that contained similarities in all three genomes. In **Table A1.1** we show that the largest category of proteins contained the Interpro domain combination IPR001128 (Cytochrome P450) and IPR002401 (E-class P450, group I). Further down the list was another set of genes with the Interpro domain combination IPR001128 (Cytochrome P450) and IPR002403 (E-class P450, group IV). While the functions of the different classes have not been determined, it is apparent that this is a key family of genes in basidiomycete biology. They

are likely involved in biomass degradation, as many of the genes are secreted. The genomic organization is similar to that of the P450 families in *P. chrysosporium* [1]□, with many genomic clusters of mixed IPR002401 (E-class P450, group I) and IPR002403 (E-class P450, group IV) domain containing genes.

In the "unique to *P. placenta*" group from **Table A1.1**, we found that a large number of Interpro categories corresponding to DNA binding were present. In addition, when we identified the largest differences in Interpro content between the various categories from the Venn diagram in **Figure A1.1**, as shown in **Table A1.2**, we noticed that categories contained DNA binding domains such as IPR001878 (Zn-finger, CCHC type), IPR007087 (Zn-finger, C2H2 type), IPR002197 (Helix-turn-helix, Fis-type), IPR001356 (Homeobox) and IPR001005 (Myb, DNA-binding). This could indicate that reprogramming of gene interaction networks is key to niche adaptation, which could explain the high number of regulatory proteins with very low similarity to those in any other characterized organism.

With the unique assembly of the *P. placenta* genome (the assembly parameters were adjusted so as to preserve the haplotypes as much as possible), we can now ask questions concerning the evolution of protein coding sequences in a manner that includes the difference in alleles. First, a caveat, we have found that not all the genes in the genome have allele pairs (see Materials and Methods, Chapter 2), the significance of which we discuss below. Briefly, we found that only about one third of the genes are represented in allele pairs. We were interested to determine if there was a pattern to genes that were found as allele pairs versus those that were not, with respect to the regions of the Venn diagram in **Figure A1.1**.

We found that the number of genes with an allele pair in the "unique to *P. placenta*" group was only 24.9%, while the number of genes in *P. placenta* with an allele pair in the "all conserved" group was a much higher 67% (p-value = 1.8 e-10). This was surprising, as we had expected that the "unique to *P. placenta*" genes would most likely include a higher number of genes with an allele pair, hypothesizing that reduced selection would result in a higher difference between haplotypes; however, the opposite seemed to be the case. While this could reflect the types of genes that were found within this group, it was not clear what could cause this dramatic difference. There are, however, two reasons why this result might be erroneous. First, it is possible that the coverage over large areas of the genome was too low (and thus our allele match would be in a gap). We note that a fair number of the assembly gaps (67%) were bounded by transposable elements, which are known to cause gaps in whole genome shotgun assemblies due to their high level of identity and wide dispersion throughout the genome [1]. This would indicate that the majority of gaps were caused by transposable elements, and not simply regions of low coverage. Second, there is the possibility, although an unlikely one, that alleles have lost so much similarity that allele identities were lower than our cutoff. To address this possibility, we performed several tests adjusting relevant parameters and found little change in the overall numbers of alleles (data not shown).

While there was a striking difference in the number of alleles per Venn set, we were also interested to find *P. placenta* alleles that showed a difference in the rate of mutations, depending on if the genes had homologs within the other two basidiomycetes in the study. Utilizing the bioperl interface [5] to the yn00 [6] algorithm, we calculated the Ka (rate of nonsynonymous nucleotide changes), Ks (rate of synonymous nucleotide changes) and Ka/Ks (ratio of nonsynonymous to synonymous changes) for all available allele pairs. As for

the above Interpro analysis, we separated the allele data into sets corresponding to the Venn diagram in **Figure A1.1**. For the sake of statistical comparisons, we assumed that genes not represented as allele pairs in the assembly did not have any difference between the haplotype copies, thus the Ka, Ks and Ka/Ks scores would simply be zero. In other words, if the nucleotide identity was 100%, alleles in the two haplotypes did not have enough difference to be separated in the assembly. We then compared the mean Ka, Ks and Ka/Ks from the regions of the Venn diagram in **Figure A1.1**. The results of this analysis are shown in **Table A1.3**.

We performed statistical analyses of the differences in the Ka, Ks and Ka/Ks Venn diagram sets of genes in order to find significant differences. While at first glance there does not seem to be large differences among the gene sets in **Table A1.3**, there is, however, a statistically significant difference in Ka (nonsynonymous substitution) between the "all conserved" group and the "unique to *P. placenta*" set (p-value = 6.815 e-8). This indicates that amino acid replacements are occurring at a significantly higher rate in the alleles from genes that were only found in *P. placenta* versus the alleles from the genes found in the all conserved group. In fact, the *P. placenta* unique genes seemed to be significantly higher than all other groups with respect to Ka, indicating relaxed or positive selection on this subset of genes. However, while the Ks (synonymous substitution rate) was higher in the "unique to *P. placenta*" category, it was not significant (p-value of > .5) when compared to other groups. Indeed, the Ka/Ks ratio was not significant, suggesting that overall rates of evolution were approximately the same among all separable allele pairs in the genome.

References

1.   Martinez D, Larrondo L, Putnam N, Gelpke M, Huang K, et al. (2004) Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. Nature Biotechnology 22: 695-700.

2.   Galagan J, Calvo S, Borkovich K, Selker E, Read N, et al. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422: 859-868.

3.   Martin F, Aerts A, Ahren D, Brun A, Danchin EGJ, et al. (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. Nature 452: 88-92.

4.   Zdobnov E, Apweiler R (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847-848.

5.   Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The Bioperl Toolkit: Perl Modules for the Life Sciences. Genome Res. 12: 1611-1618. A

6.   Yang Z, Nielsen R (2000) Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. Mol Biol Evol 17: 32-43.
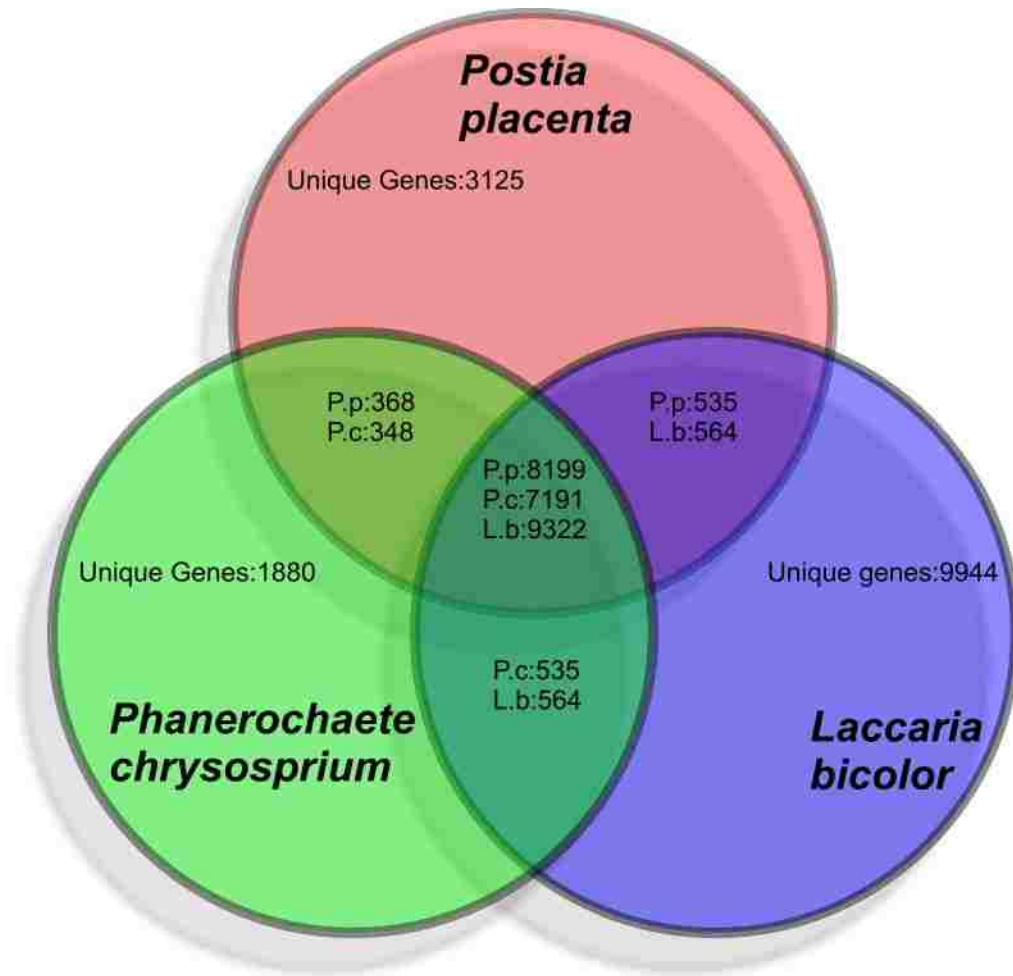
Figure A1.1. Venn diagram showing the separation of proteins from three basidiomycetes (*Postia placenta, Laccaria bicolor* and *Phanerochaete chrysosporium*) into four overlapping and three non-overlapping areas. Similarity was calculated using BLAST, allowing for a bit score of 50 or more and a coverage (alignment length divided by original sequence length) of twenty percent.

Table A1.1. Top forty most abundant Interpro domain combinations from the *P. placenta* genomeeparated as in Figure A1.1. In all tables, the first column is the Interpro annotation, while the second column is the number of proteins with that combination.

Interpro domain/domain combination	number of genes in *P. placenta*
Unique to P. placenta group	
IPR001878::Zn-finger, CCHC type	246
IPR007087::Zn-finger, C2H2 type	96
IPR001810::Cyclin-like F-box	35
IPR000345::Cytochrome c heme-binding site	9
IPR000772::Ricin B lectin	8
IPR006209::EGF-like	8
IPR007087::Zn-finger, C2H2 type,IPR000345::Cytochrome c heme-binding site	7
IPR002086::Aldehyde dehydrogenase	5
IPR002110::Ankyrin	5
IPR001395::Aldo/keto reductase	4
IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	4
IPR002048::Calcium-binding EF-hand	4
IPR002197::Helix-turn-helix, Fis-type	4
IPR001356::Homeobox	3
IPR003006::Immunoglobulin/major histocompatibility complex	3
IPR010916::TonB box, N-terminal	3
IPR001005::Myb, DNA-binding	3
IPR001901::Protein secE/sec61-gamma protein	3
IPR000585::Hemopexin	3
IPR001680::G-protein beta WD-40 repeat	3
IPR000169::Peptidase, eukaryotic cysteine peptidase active site	3
IPR001052::Rubredoxin,IPR007087::Zn-finger, C2H2 type	3
IPR000719::Protein kinase	3
IPR011118::Tannase and feruloyl esterase	3
IPR001810::Cyclin-like F-box,IPR006162::Phosphopantetheine attachment site	2
IPR001412::Aminoacyl-tRNA synthetase, class I	2
IPR000209::Peptidase S8 and S53, subtilisin, kexin, sedolisin	2
IPR001087::Lipolytic enzyme, G-D-S-L	2
IPR001540::Glycoside hydrolase, family 20	2
IPR007065::HPP	2
IPR000910::HMG1/2 (high mobility group) box	2
IPR000253::Forkhead-associated	2
IPR001199::Cytochrome b5	2
IPR000408::Regulator of chromosome condensation, RCC1	2
IPR000566::Lipocalin-related protein and Bos/Can/Equ allergen	2
IPR000048::IQ calmodulin-binding region	2
IPR001862::Membrane attack complex component/perforin/complement C9	2
IPR002150::Ribosomal protein L31	2
IPR000276::Rhodopsin-like GPCR superfamily	2
IPR000194::H+ transporting two-sector ATPase, alpha/beta subunit, central region	2

All conserved group	
IPR001128::Cytochrome P450,IPR002401::E-class P450, group I	139
IPR002347::Glucose/ribitol dehydrogenase,IPR002198::Short-chain dehydrogenase/reductase SDR	90
IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase,IPR001245::Tyrosine protein kinase,IPR008271::Serine/threonine protein kinase, active site	79
IPR001680::G-protein beta WD-40 repeat	77
IPR001810::Cyclin-like F-box	62
IPR007114::Major facilitator superfamily	60
IPR001395::Aldo/keto reductase	59
IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	48
IPR008266::Tyrosine protein kinase, active site	47
IPR001878::Zn-finger, CCHC type	41
IPR000210::BTB/POZ	39
IPR002085::Zinc-containing alcohol dehydrogenase superfamily	38
IPR007087::Zn-finger, C2H2 type	35
IPR000379::Esterase/lipase/thioesterase	31
IPR001461::Peptidase A1, pepsin,IPR001969::Peptidase aspartic, active site	31
IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase	28
IPR000172::Glucose-methanol-choline oxidoreductase,IPR007867::GMC oxidoreductase	25
IPR001440::TPR repeat	25
IPR001248::Permease for cytosine/purines, uracil, thiamine, allantoin	23
IPR000608::Ubiquitin-conjugating enzymes	22
IPR000719::Protein kinase	22
IPR001128::Cytochrome P450,IPR002403::E-class P450, group IV	21
IPR002110::Ankyrin	20
IPR001138::Fungal transcriptional regulatory protein, N-terminal	20
IPR002086::Aldehyde dehydrogenase	19
IPR000910::HMG1/2 (high mobility group) box	19
IPR001138::Fungal transcriptional regulatory protein, N-terminal,IPR007219::Fungal specific transcription factor	18
IPR005829::Sugar transporter superfamily,IPR007114::Major facilitator superfamily,IPR003663::Sugar transporter,IPR005828::General substrate transporter	18
IPR000873::AMP-dependent synthetase and ligase	18
IPR000182::GCN5-related N-acetyltransferase	17
IPR001155::NADH:flavin oxidoreductase/NADH oxidase	17
IPR001841::Zn-finger, RING	17
IPR003593::AAA ATPase,IPR003439::ABC transporter,IPR001687::ATP/GTP-binding site motif A (P-loop),IPR001140::ABC transporter, transmembrane region	17
IPR000051::SAM (and some other nucleotide) binding motif	17
IPR011545::DEAD/DEAH box helicase, N-terminal,IPR000629::ATP-dependent helicase, DEAD-box,IPR001410::DEAD/DEAH box helicase,IPR001650::Helicase, C-terminal	15
IPR007114::Major facilitator superfamily,IPR005828::General substrate transporter	15
IPR001993::Mitochondrial substrate carrier	15
IPR004046::Glutathione S-transferase, C-terminal,IPR004045::Glutathione S-transferase, N-terminal15	
IPR007269::Isoprenylcysteine carboxyl methyltransferase	14
IPR001547::Glycoside hydrolase, family 5	14

P. placenta and Laccaria bicolor conserved only group	
IPR003812::Filamentation induced by cAMP protein Fic	6
IPR002048::Calcium-binding EF-hand	5
IPR007087::Zn-finger, C2H2 type	4
IPR001841::Zn-finger, RING	4
IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	3
IPR001163::Small nuclear ribonucleoprotein (Sm protein),IPR006649::snRNP	3
IPR001005::Myb, DNA-binding	3
IPR001810::Cyclin-like F-box	3
IPR010721::Protein of unknown function DUF1295	2
IPR000834::Peptidase M14, carboxypeptidase A	2
IPR006785::Peroxisomal membrane anchor protein (Pex14p)	2
IPR006683::Thioesterase superfamily	2
IPR007720::N-acetylglucosaminyl transferase component	2
IPR008576::Eukaryotic protein of unknown function DUF858	2
IPR005834::Haloacid dehalogenase-like hydrolase,IPR006402::HAD-superfamily hydrolase, subfamily IA, variant 3	2
IPR008011::Complex 1 LYR protein	2
IPR001705::Ribosomal protein L33	2
IPR002610::Rhomboid-like protein	2
IPR002893::Zn-finger, MYND type	2
IPR002052::N-6 Adenine-specific DNA methylase	2
IPR001356::Homeobox	1
IPR005607::BSD	1
IPR003511::DNA-binding HORMA	1
IPR002677::Ribosomal L32p protein	1
IPR003591::Leucine-rich repeat, typical subtype,IPR001611::Leucine-rich repeat	1
IPR000352::Class I peptide chain release factor	1
IPR000529::Ribosomal protein S6	1
IPR007745::Cytochrome C oxidase copper chaperone	1
IPR003521::Nucleotide-sensitive chloride conductance regulator	1
IPR005599::Alg9-like mannosyltransferase	1
IPR000684::Eukaryotic RNA polymerase II heptapeptide repeat,IPR006845::Pex2 / Pex12, N-terminal	1
IPR002833::Protein of unknown function UPF0099	1
IPR005834::Haloacid dehalogenase-like hydrolase,IPR000319::Aspartate-semialdehyde dehydrogenase,IPR006402::HAD-superfamily hydrolase, subfamily IA, variant 3	1
IPR007305::Got1-like protein	1
IPR006652::Kelch repeat,IPR011498::Kelch	1
IPR006809::TAFII28-like protein	1
IPR001440::TPR repeat	1
IPR004942::Roadblock/LC7	1
IPR003103::Apoptosis regulator Bcl-2 protein, BAG,IPR000626::Ubiquitin	1
IPR003610::Carbohydrate-binding domain, family V/XII,IPR006616::Protein of unknown function DM9	1

P. placenta and P. chrysosporium conserved only group	
IPR003866::Isoflavone reductase	6
IPR001810::Cyclin-like F-box	5
IPR004304::Acetamidase/Formamidase	4
IPR000433::Zn-finger, ZZ type	4
IPR002509::Polysaccharide deacetylase	3
IPR000292::Formate/nitrite transporter	3
IPR005833::Haloacid dehalogenase/epoxide hydrolase,IPR005834::Haloacid dehalogenase-like hydrolase,IPR006388::HAD-superfamily hydrolase, subfamily IA, variant 2	3
IPR001466::Beta-lactamase	3
IPR000379::Esterase/lipase/thioesterase,IPR010497::Epoxide hydrolase, N-terminal,IPR000639::Epoxide hydrolase	3
IPR001000::Glycoside hydrolase, family 10	2
IPR000630::Ribosomal protein S8,IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	2
IPR011021::Arrestin, N-terminal	2
IPR001354::Mandelate racemase/muconate lactonizing enzyme	2
IPR001878::Zn-finger, CCHC type	2
IPR002110::Ankyrin	2
IPR002642::Lysophospholipase, catalytic region	2
IPR004827::Basic-leucine zipper (bZIP) transcription factor	2
IPR000209::Peptidase S8 and S53, subtilisin, kexin, sedolisin,IPR002110::Ankyrin	2
IPR005000::HpcH/HpaI aldolase	2
IPR001841::Zn-finger, RING	2
IPR000834::Peptidase M14, carboxypeptidase A	2
IPR000816::Peptidase C15, pyroglutamyl peptidase I	2
IPR007113::Cupin region	2
IPR001952::Alkaline phosphatase	2
IPR005593::D-xylulose 5-phosphate/D-fructose 6-phosphate phosphoketolase	2
IPR001360::Glycoside hydrolase, family 1	2
IPR000194::H+-transporting two-sector ATPase, alpha/beta subunit, central region	2
IPR002828::Survival protein SurE	2
IPR001951::Histone H4	1
IPR007781::Alpha-N-acetylglucosaminidase,IPR001547::Glycoside hydrolase, family 5	1
IPR007229::Nicotinate phosphoribosyltransferase and related,IPR006406::Nicotinate phosphoribosyltransferase	1
IPR007318::Phospholipid methyltransferase	1
IPR000379::Esterase/lipase/thioesterase	1
IPR000374::Phosphatidate cytidylyltransferase	1
IPR010720::Alpha-L-arabinofuranosidase, C-terminal	1
IPR001345::Phosphoglycerate/bisphosphoglycerate mutase	1
IPR006050::DNA photolyase, N-terminal,IPR005101::DNA photolyase, FAD-binding,IPR006051::DNA photolyase, FAD- binding N-terminal	1
IPR004547::Glucosamine-6-phosphate isomerase,IPR006148::Glucosamine/galactosamine-6-phosphate isomerase	1
IPR001327::FAD-dependent pyridine nucleotide-disulphide oxidoreductase,IPR002922::Thiamine biosynthesis Thi4 protein	1
IPR005162::Retrotransposon gag protein,IPR001878::Zn-finger, CCHC type	1

Table A1.2. Top 20 Interpro domain differences and top 10 unique Interpro domains in P. placenta predicted genes split into sets according to the Venn diagram in Figure A1.1. The abundance numerical columns (columns two and four throughout) show the ratio of the number of genes with that particular domain or domain combination divided by the total number of Interpro results for that particular Venn split set.

Unique to P. placenta Vs. All Conserved			
Greater abundance in "Unique to P. placenta"	Ratio of abundance	Greater abundance in "all conserved"	Ratio of abundance
IPR001878::Zn-finger, CCHC type	0.37334	IPR007114::Major facilitator superfamily	0.00938
IPR007087::Zn-finger, C2H2 type	0.14224	IPR001680::G-protein beta WD-40 repeat	0.00937
IPR001810::Cyclin-like F-box	0.04289	IPR001395::Aldo/keto reductase	0.00455
IPR000345::Cytochrome c heme-binding site	0.01157	IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase	0.00355
IPR006209::EGF-like	0.01111	IPR001248::Permease for cytosine/purines, uracil, thiamine, allantoin	0.00264
IPR007087::Zn-finger, C2H2 type,IPR000345::Cytochrome c heme-binding site	0.01047	IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	0.00255
IPR002197::Helix-turn-helix, Fis-type	0.00601	IPR001138::Fungal transcriptional regulatory protein, N-terminal	0.00209
IPR000585::Hemopexin	0.00446	IPR000873::AMP-dependent synthetase and ligase	0.00173
IPR002086::Aldehyde dehydrogenase	0.00428	IPR000051::SAM (and some other nucleotide) binding motif	0.00155
IPR010916::TonB box, N-terminal	0.00428	IPR001547::Glycoside hydrolase, family 5	0.00100
IPR002110::Ankyrin	0.00410	IPR005123::2OG-Fe(II) oxygenase superfamily	0.00064
IPR003006::Immunoglobulin/major histocompatibility complex	0.00410	IPR006025::Peptidase M, neutral zinc metallopeptidases, zinc-binding site	0.00045
IPR002048::Calcium-binding EF-hand	0.00401	IPR000910::HMG1/2 (high mobility group) box	0.00036
IPR001356::Homeobox	0.00392		
IPR000169::Peptidase, eukaryotic cysteine peptidase active site	0.00373		
IPR002035::von Willebrand factor, type A	0.00291		
IPR002016::Haem peroxidase, plant/fungal/bacterial	0.00291		
IPR001005::Myb, DNA-binding	0.00282		
IPR000432::DNA mismatch repair protein MutS, C-terminal	0.00273		
IPR001087::Lipolytic enzyme, G-D-S-L	0.00273		

Unique to *P. placenta* Vs. All Conserved			
Greater abundance in "Unique to *P. placenta*"	**Ratio of abundance**	**Greater abundance in "all conserved"**	**Ratio of abundance**
IPR001878::Zn-finger, CCHC type	0.37334	IPR007114::Major facilitator superfamily	0.00938
IPR007087::Zn-finger, C2H2 type	0.14224	IPR001680::G-protein beta WD-40 repeat	0.00937
IPR001810::Cyclin-like F-box	0.04289	IPR001395::Aldo/keto reductase	0.00455
IPR000345::Cytochrome c heme-binding site	0.01157	IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase	0.00355
IPR006209::EGF-like	0.01111	IPR001248::Permease for cytosine/purines, uracil, thiamine, allantoin	0.00264
IPR007087::Zn-finger, C2H2 type,IPR000345::Cytochrome c heme-binding site	0.01047	IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	0.00255
IPR002197::Helix-turn-helix, Fis-type	0.00601	IPR001138::Fungal transcriptional regulatory protein, N-terminal	0.00209
IPR000585::Hemopexin	0.00446	IPR000873::AMP-dependent synthetase and ligase	0.00173
IPR002086::Aldehyde dehydrogenase	0.00428	IPR000051::SAM (and some other nucleotide) binding motif	0.00155
IPR010916::TonB box, N-terminal	0.00428	IPR001547::Glycoside hydrolase, family 5	0.00100
IPR002110::Ankyrin	0.00410	IPR005123::2OG-Fe(II) oxygenase superfamily	0.00064
IPR003006::Immunoglobulin/major histocompatibility complex	0.00410	IPR006025::Peptidase M, neutral zinc metallopeptidases, zinc-binding site	0.00045
IPR002048::Calcium-binding EF-hand	0.00401	IPR000910::HMG1/2 (high mobility group) box	0.00036
IPR001356::Homeobox	0.00392		
IPR000169::Peptidase, eukaryotic cysteine peptidase active site	0.00373		
IPR002035::von Willebrand factor, type A	0.00291		
IPR002016::Haem peroxidase, plant/fungal/bacterial	0.00291		
IPR001005::Myb, DNA-binding	0.00282		
IPR000432::DNA mismatch repair protein MutS, C-terminal	0.00273		
IPR001087::Lipolytic enzyme, G-D-S-L	0.00273		

unique to *P. placenta* VS *P. placenta - L. bicolor* group			
Greater abundance in "Unique to *P. placenta*"	**Ratio of abundance**	**Greater abundance in "*P. placenta – L. bicolor* conserved only"**	**Ratio of abundance**
IPR007087::Zn-finger, C2H2 type	0.12900	IPR002048::Calcium-binding EF-hand	0.01832
IPR001810::Cyclin-like F-box	0.03947	IPR001841::Zn-finger, RING	0.01651
IPR002086::Aldehyde dehydrogenase	0.00284	IPR001005::Myb, DNA-binding	0.01006
IPR002110::Ankyrin	0.00284	IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	0.00851
		IPR002052::N-6 Adenine-specific DNA methylase	0.00826
		IPR000834::Peptidase M14, carboxypeptidase A	0.00826
		IPR001965::Zn-finger-like, PHD finger	0.00335
		IPR005829::Sugar transporter superfamily	0.00335
		IPR003812::Filamentation induced by cAMP protein Fic,IPR007087::Zn-finger, C2H2 type	0.00335
		IPR001138::Fungal transcriptional regulatory protein, N-terminal	0.00335
		IPR001202::WW/Rsp5/WWP	0.00335
		IPR006652::Kelch repeat,IPR011498::Kelch	0.00335
		IPR000910::HMG1/2 (high mobility group) box	0.00181
		IPR000276::Rhodopsin-like GPCR superfamily	0.00181
		IPR001159::Double-stranded RNA binding	0.00181
		IPR000209::Peptidase S8 and S53, subtilisin, kexin, sedolisin	0.00181
		IPR001356::Homeobox	0.00026
		IPR001680::G-protein beta WD-40 repeat	0.00026

Unique to *P. placenta* Vs *P. placenta* - *P. chrysosporium*			
Unique Domains in "unique to *P. placenta*"	Ratio of abundance	Unique domains in "*P. placenta* – *P. chrysosporium* conserved only"	Ratio of abundance
IPR007087::Zn-finger, C2H2 type	0.14861	IPR004304::Acetamidase/Formamidase	0.02581
IPR000345::Cytochrome c heme-binding site	0.01393	IPR000433::Zn-finger, ZZ type	0.02581
IPR000772::Ricin B lectin	0.01238	IPR002509::Polysaccharide deacetylase	0.01935
IPR006209::EGF-like	0.01238	IPR000292::Formate/nitrite transporter	0.01935
IPR007087::Zn-finger, C2H2 type,IPR000345::Cytochrome c heme-binding site	0.01084	IPR005833::Haloacid dehalogenase/epoxide hydrolase,IPR005834::Haloacid dehalogenase-like hydrolase,IPR006388::HAD-superfamily hydrolase, subfamily IA, variant 2	0.01935
IPR002086::Aldehyde dehydrogenase	0.00774	IPR001466::Beta-lactamase	0.01935
IPR001395::Aldo/keto reductase	0.00619	IPR000379::Esterase/lipase/thioesterase,IPR010497:: Epoxide hydrolase, N-terminal,IPR000639::Epoxide hydrolase0.0193548387096774	
IPR002197::Helix-turn-helix, Fis-type	0.00619	IPR001000::Glycoside hydrolase, family 10	0.01290
IPR001356::Homeobox	0.00464	IPR000630::Ribosomal protein S8,IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	0.01290
IPR003006::Immunoglobulin/major histocompatibility complex	0.00464	IPR011021::Arrestin, N-terminal	0.01290
IPR010916::TonB box, N-terminal	0.00464	IPR001354::Mandelate racemase/muconate lactonizing enzyme	0.01290
IPR001005::Myb, DNA-binding	0.00464	IPR002642::Lysophospholipase, catalytic region	0.01290
IPR001901::Protein secE/sec61-gamma protein	0.00464	IPR000209::Peptidase S8 and S53, subtilisin, kexin, sedolisin,IPR002110::Ankyrin	0.01290
IPR000585::Hemopexin	0.00464	IPR005000::HpcH/HpaI aldolase	0.01290
IPR001680::G-protein beta WD-40 repeat	0.00464	IPR000816::Peptidase C15, pyroglutamyl peptidase I	0.01290
IPR000169::Peptidase, eukaryotic cysteine peptidase active site	0.00464	IPR007113::Cupin region	0.01290
IPR001052::Rubredoxin,IPR007087::Zn-finger, C2H2 type	0.00464	IPR001952::Alkaline phosphatase	0.01290
IPR000719::Protein kinase	0.00464	IPR005593::D-xylulose 5-phosphate/D-fructose 6-phosphate phosphoketolase	0.01290
IPR011118::Tannase and feruloyl esterase	0.00464	IPR001360::Glycoside hydrolase, family 1	0.01290
IPR001810::Cyclin-like F-box,IPR006162::Phosphopantetheine attachment site	0.00310	IPR002828::Survival protein SurE	0.01290

Unique to *P. placenta* Vs *P. placenta - P. chrysosporium*			
Greater abundance in "Unique to *P. placenta*"	**Ratio of abundance**	**Greater abundance in "*P. placenta – P. chrysosporium* conserved only"**	**Ratio of abundance**
IPR001878::Zn-finger, CCHC type	0.36790	IPR003866::Isoflavone reductase	0.03716
IPR001810::Cyclin-like F-box	0.02192	IPR004827::Basic-leucine zipper (bZIP) transcription factor	0.01136
		IPR000834::Peptidase M14, carboxypeptidase A	0.01136
		IPR000194::H+-transporting two-sector ATPase, alpha/beta subunit, central region	0.00981
		IPR001841::Zn-finger, RING	0.00981
		IPR002110::Ankyrin	0.00516
		IPR000051::SAM (and some other nucleotide) binding motif	0.00490
		IPR002052::N-6 Adenine-specific DNA methylase	0.00490
		IPR002129::Pyridoxal-dependent decarboxylase	0.00490
		IPR001345::Phosphoglycerate/bisphosphoglycerate mutase	0.00490
		IPR001487::Bromodomain	0.00490
		IPR005162::Retrotransposon gag protein	0.00490
		IPR001810::Cyclin-like F-box,IPR000577::Carbohydrate kinase, FGGY	0.00490
		IPR002048::Calcium-binding EF-hand	0.00026
		IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	0.00026

unique to *P. placenta* VS *P. placenta - L. bicolor* group			
Unique Domains in "unique to *P. placenta*"	**Ratio of abundance**	**Unique domains in "*P. placenta – L. bicolor* conserved only"**	**Ratio of abundance**
IPR001878::Zn-finger, CCHC type	0.38080	IPR003812::Filamentation induced by cAMP protein Fic	0.02941
IPR000345::Cytochrome c heme-binding site	0.01393	IPR001163::Small nuclear ribonucleoprotein (Sm protein),IPR006649::snRNP	0.01471
IPR000772::Ricin B lectin	0.01238	IPR010721::Protein of unknown function DUF1295	0.00980
IPR006209::EGF-like	0.01238	IPR006785::Peroxisomal membrane anchor protein (Pex14p)	0.00980
IPR007087::Zn-finger, C2H2 type,IPR000345::Cytochrome c heme-binding site	0.01084	IPR006683::Thioesterase superfamily	0.00980
IPR001395::Aldo/keto reductase	0.00619	IPR007720::N-acetylglucosaminyl transferase component	0.00980
IPR002197::Helix-turn-helix, Fis-type	0.00619	IPR008576::Eukaryotic protein of unknown function DUF858	0.00980
IPR003006::Immunoglobulin/major histocompatibility complex	0.00464	IPR005834::Haloacid dehalogenase-like hydrolase,IPR006402::HAD-superfamily hydrolase, subfamily IA, variant 30.009803921568627445	
IPR010916::TonB box, N-terminal	0.00464	IPR008011::Complex 1 LYR protein	0.00980
IPR001901::Protein secE/sec61-gamma protein	0.00464	IPR001705::Ribosomal protein L33	0.00980
IPR000585::Hemopexin	0.00464	IPR002610::Rhomboid-like protein	0.00980
IPR000169::Peptidase, eukaryotic cysteine peptidase active site	0.00464	IPR002893::Zn-finger, MYND type	0.00980
IPR001052::Rubredoxin,IPR007087::Zn-finger, C2H2 type	0.00464	IPR005607::BSD	0.00490

All conserved VS *P. placenta* - *L. bicolor* group			
Greater abundance in "All conserved"	**Ratio of abundance**	**Greater abundance in "P. placenta – L. bicolor conserved only"**	**Ratio of abundance**
IPR001680::G-protein beta WD-40 repeat	0.00912	IPR002048::Calcium-binding EF-hand	0.02232
IPR000210::BTB/POZ	0.00220	IPR001841::Zn-finger, RING	0.01651
IPR000379::Esterase/lipase/thioesterase	0.00074	IPR007087::Zn-finger, C2H2 type	0.01323
		IPR001005::Myb, DNA-binding	0.01289
		IPR001163::Small nuclear ribonucleoprotein (Sm protein),IPR006649::snRNP	0.01270
		IPR006683::Thioesterase superfamily	0.00962
		IPR002610::Rhomboid-like protein	0.00962
		IPR005834::Haloacid dehalogenase-like hydrolase,IPR006402::HAD-superfamily hydrolase, subfamily IA, variant 3	0.00944
		IPR008011::Complex 1 LYR protein	0.00944
		IPR000834::Peptidase M14, carboxypeptidase A	0.00926
		IPR002052::N-6 Adenine-specific DNA methylase	0.00908
		IPR002893::Zn-finger, MYND type	0.00817
		IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	0.00597
		IPR001544::Aminotransferase, class IV	0.00472
		IPR000352::Class I peptide chain release factor	0.00472
		IPR008278::4'-phosphopantetheinyl transferase	0.00472
		IPR006677::tRNA intron endonuclease, catalytic C-terminal	0.00472
		IPR002156::RNase H	0.00472
		IPR000754::Ribosomal protein S9	0.00472
		IPR001965::Zn-finger-like, PHD finger	0.00472

All conserved VS *P. placenta* - *L. bicolor* group			
Unique Domains in "All conserved"	**Ratio of abundance**	**Unique domains in "*P. placenta* – *L. bicolor* conserved only"**	**Ratio of abundance**
IPR001128::Cytochrome P450,IPR002401::E-class P450, group I	0.02531	IPR003812::Filamentation induced by cAMP protein Fic	0.02941
IPR002347::Glucose/ribitol dehydrogenase,IPR002198::Short-chain dehydrogenase/reductase SDR	0.01639	IPR010721::Protein of unknown function DUF1295	0.00980
IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase,IPR001245::Tyrosine protein kinase,IPR008271::Serine/threonine protein kinase, active site	0.01438	IPR006785::Peroxisomal membrane anchor protein (Pex14p)	0.00980
IPR007114::Major facilitator superfamily	0.01092	IPR007720::N-acetylglucosaminyl transferase component	0.00980
IPR001395::Aldo/keto reductase	0.01074	IPR008576::Eukaryotic protein of unknown function DUF858	0.00980
IPR008266::Tyrosine protein kinase, active site	0.00856	IPR001705::Ribosomal protein L33	0.00980
IPR001878::Zn-finger, CCHC type	0.00747	IPR005607::BSD	0.00490
IPR002085::Zinc-containing alcohol dehydrogenase superfamily	0.00692	IPR002677::Ribosomal L32p protein	0.00490
IPR001461::Peptidase A1, pepsin,IPR001969::Peptidase aspartic, active site	0.00564	IPR007745::Cytochrome C oxidase copper chaperone	0.00490
IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase	0.00510	IPR000529::Ribosomal protein S6	0.00490
IPR000172::Glucose-methanol-choline oxidoreductase,IPR007867::GMC oxidoreductase	0.00455	IPR003521::Nucleotide-sensitive chloride conductance regulator	0.00490
IPR001248::Permease for cytosine/purines, uracil, thiamine, allantoin	0.00419	IPR002833::Protein of unknown function UPF0099	0.00490
IPR000608::Ubiquitin-conjugating enzymes	0.00401	IPR000684::Eukaryotic RNA polymerase II heptapeptide repeat,IPR006845::Pex2 / Pex12, N-terminal	0.00490
IPR000719::Protein kinase	0.00401	IPR005834::Haloacid dehalogenase-like hydrolase,IPR000319::Aspartate-semialdehyde dehydrogenase,IPR006402::HAD-superfamily hydrolase, subfamily IA, variant 3	0.00490
IPR001128::Cytochrome P450,IPR002403::E-class P450, group IV	0.00382	IPR007305::Got1-like protein	0.00490
IPR001138::Fungal transcriptional regulatory protein, N-terminal,IPR007219::Fungal specific transcription factor	0.00328	IPR006809::TAFII28-like protein	0.00490
IPR005829::Sugar transporter superfamily,IPR007114::Major facilitator superfamily,IPR003663::Sugar transporter,IPR005828::General substrate transporter	0.00328	IPR003103::Apoptosis regulator Bcl-2 protein, BAG,IPR000626::Ubiquitin	0.00490
IPR000873::AMP-dependent synthetase and ligase	0.00328	IPR004942::Roadblock/LC7	0.00490
IPR001155::NADH:flavin oxidoreductase/NADH oxidase	0.00310	IPR003610::Carbohydrate-binding domain, family V/XII,IPR006616::Protein of unknown function DM9	0.00490
IPR003593::AAA ATPase,IPR003439::ABC transporter,IPR001687::ATP/GTP-binding site motif A (P-loop),IPR001140::ABC transporter, transmembrane region	0.00310	IPR007277::Protein of unknown function DUF396	0.00490

chrysporium			
Unique Domains in "All conserved"	**abundance**	**unique domains in "*P. placenta – P. chrysosporium conserved only*"**	**ratio of abundance**
IPR001128::Cytochrome P450,IPR002401::E-class P450, group I	0.02531	IPR003866::Isoflavone reductase	0.03871
IPR002347::Glucose/ribitol dehydrogenase,IPR002198::Short-chain dehydrogenase/reductase SDR	0.01639	IPR004304::Acetamidase/Formamidase	0.02581
IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase,IPR001245::Tyrosine protein kinase,IPR008271::Serine/threonine protein kinase, active site	0.01438	IPR000433::Zn-finger, ZZ type	0.02581
IPR001680::G-protein beta WD-40 repeat	0.01402	IPR000292::Formate/nitrite transporter	0.01935
IPR007114::Major facilitator superfamily	0.01092	IPR000379::Esterase/lipase/thioesterase,IPR010497::Epoxide hydrolase, N-terminal,IPR000639::Epoxide hydrolase	0.0193548387096774
IPR001395::Aldo/keto reductase	0.01074	IPR001000::Glycoside hydrolase, family 10	0.01290
IPR008266::Tyrosine protein kinase, active site	0.00856	IPR000630::Ribosomal protein S8,IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	0.01290
IPR000210::BTB/POZ	0.00710	IPR011021::Arrestin, N-terminal	0.01290
IPR002085::Zinc-containing alcohol dehydrogenase superfamily	0.00692	IPR001354::Mandelate racemase/muconate lactonizing enzyme	0.01290
IPR007087::Zn-finger, C2H2 type	0.00637	IPR002642::Lysophospholipase, catalytic region	0.01290
IPR001461::Peptidase A1, pepsin,IPR001969::Peptidase aspartic, active site	0.00564	IPR000209::Peptidase S8 and S53, subtilisin, kexin, sedolisin,IPR002110::Ankyrin	0.01290
IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase	0.00510	IPR005000::HpcH/HpaI aldolase	0.01290
IPR000172::Glucose-methanol-choline oxidoreductase,IPR007867::GMC oxidoreductase	0.00455	IPR000816::Peptidase C15, pyroglutamyl peptidase I	0.01290
IPR001440::TPR repeat	0.00455	IPR007113::Cupin region	0.01290
IPR001248::Permease for cytosine/purines, uracil, thiamine, allantoin	0.00419	IPR001952::Alkaline phosphatase	0.01290
IPR000608::Ubiquitin-conjugating enzymes	0.00401	IPR005593::D-xylulose 5-phosphate/D-fructose 6-phosphate phosphoketolase	0.01290
IPR000719::Protein kinase	0.00401	IPR001360::Glycoside hydrolase, family 1	0.01290
IPR001128::Cytochrome P450,IPR002403::E-class P450, group IV	0.00382	IPR001951::Histone H4	0.00645
IPR001138::Fungal transcriptional regulatory protein, N-terminal	0.00364	IPR007781::Alpha-N-acetylglucosaminidase,IPR001547::Glycoside hydrolase, family 5	0.00645
IPR002086::Aldehyde dehydrogenase	0.00346	IPR007229::Nicotinate phosphoribosyltransferase and related,IPR006406::Nicotinate phosphoribosyltransferase	0.00645

P. placenta - Pchrysosporium group Vs. P. placenta - L. bicolor group		
Greater abundance in " P. placenta - L. bicolor conserved only"	ratio of abundance	
IPR002048::Calcium-binding EF-hand	0.01802	There are no groups Larger in P.placenta - P.chrysosporium
IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	0.00821	
IPR001841::Zn-finger, RING	0.00662	
IPR002052::N-6 Adenine-specific DNA methylase	0.00331	
IPR001810::Cyclin-like F-box	0.01776	
IPR002110::Ankyrin	0.00809	
IPR000834::Peptidase M14, carboxypeptidase A	0.00318	
IPR000379::Esterase/lipase/thioesterase	0.00159	
IPR000909::Phosphatidylinositol-specific phospholipase C, X region	0.00159	
IPR000182::GCN5-related N-acetyltransferase	0.00159	

P. placenta - Pchrysosporium group Vs. P. placenta - L. bicolor group			
Unique Domains in "P. placenta – L. bicolor conserved only"	abundance	Unique domains in "P. placenta – P. chrysosporium conserved only"	ratio of abundance
IPR003812::Filamentation induced by cAMP protein Fic	0.02941	IPR003866::Isoflavone reductase	0.03896
IPR007087::Zn-finger, C2H2 type	0.01961	IPR004304::Acetamidase/Formamidase	0.02597
IPR001163::Small nuclear ribonucleoprotein (Sm protein),IPR006649::snRNP	0.01471	IPR000433::Zn-finger, ZZ type	0.02597
IPR001005::Myb, DNA-binding	0.01471	IPR002509::Polysaccharide deacetylase	0.01948
IPR010721::Protein of unknown function DUF1295	0.00980	IPR000292::Formate/nitrite transporter	0.01948
IPR006785::Peroxisomal membrane anchor protein (Pex14p)	0.00980	IPR005833::Haloacid dehalogenase/epoxide hydrolase,IPR005834::Haloacid dehalogenase-like hydrolase,IPR006388::HAD-superfamily hydrolase, subfamily IA, variant 2	0.01948
IPR006683::Thioesterase superfamily	0.00980	IPR001466::Beta-lactamase	0.01948
IPR007720::N-acetylglucosaminyl transferase component	0.00980	IPR000379::Esterase/lipase/thioesterase,IPR010497::Epoxide hydrolase, N-terminal,IPR000639::Epoxide hydrolase0.0194805194805195	
IPR008576::Eukaryotic protein of unknown function DUF858	0.00980	IPR001000::Glycoside hydrolase, family 10	0.01299
IPR005834::Haloacid dehalogenase-like hydrolase,IPR006402::HAD-superfamily hydrolase, subfamily IA, variant 3	0.00980	IPR000630::Ribosomal protein S8,IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	0.01299
IPR008011::Complex 1 LYR protein	0.00980	IPR011021::Arrestin, N-terminal	0.01299
IPR001705::Ribosomal protein L33	0.00980	enzyme	0.01299
IPR002610::Rhomboid-like protein	0.00980	IPR001878::Zn-finger, CCHC type	0.01299
IPR002893::Zn-finger, MYND type	0.00980	IPR002642::Lysophospholipase, catalytic region	0.01299
IPR005607::BSD	0.00490	IPR004827::Basic-leucine zipper (bZIP) transcription factor	0.01299
IPR001356::Homeobox	0.00490	IPR000209::Peptidase S8 and S53, subtilisin, kexin, sedolisin,IPR002110::Ankyrin	0.01299
IPR002677::Ribosomal L32p protein	0.00490	IPR005000::HpcH/Hpal aldolase	0.01299
IPR003511::DNA-binding HORMA	0.00490	IPR000816::Peptidase C15, pyroglutamyl peptidase I	0.01299
IPR003591::Leucine-rich repeat, typical subtype,IPR001611::Leucine-rich repeat	0.00490	IPR007113::Cupin region	0.01299
IPR000352::Class I peptide chain release factor	0.00490	IPR001952::Alkaline phosphatase	0.01299

Table A1.3. Mean Ka, Ks and Ka/Ks for alleles within the split sets from the Venn diagram in Figure A1.1.

Set	Ka	Ks	Ka/Ks	Protein % id	Nucleotide % id
Unique to *P. placenta*	0.04778	0.08679	0.5265	95.69	96.66
Conserved in all three genomes	0.02841	0.0772	0.5544	96.98	97.31
Conserved only between *P. placenta* and *L. bicolor*	0.03901	0.08925	1.082	96.39	97.05
Conserved only between *P. placenta* and *P. chrysosporium*	0.0364	0.08444	0.9608	96.31	96.92
Overall	0.03406	0.0804	0.5826	96.6	97.12

ADDENDUM 2:

COMPARISON OF SYNTENY BETWEEN POSTIA PLACENTA

AND RELATED BASIDIOMYCETES

The dynamic nature of genomes has become evident with progression of the post-genomic era. Alterations in the organization of the genome, caused by genes moving within the chromosome[1] or to a different chromosome[2] are now well established but poorly understood events. Researchers studying eukaryotic genomes have long wondered if this movement of genes is random or if selection is at work. Investigations in the last decade have determined that repetitive genomic elements seem to play a large role[3], as well as genes that are lineage specific[4]. However, few whole genome studies that tie gene function to alterations in genomic organization have been conducted.

To understand how the architecture of *Postia placenta* has changed, we calculated the remaining amount of synteny (groups of genes remaining on the same chromosome since the last common ancestor of the genomes in comparison) between *Postia placenta* and two basidiomycetes, *Phanerochaete chrysosporium[5]* and *Laccaria bicolor[6]* To perform this test, we used an ortholog anchored version of the Maximum Gap Cluster technique[7] as implemented in the recent Martinez et al. study[4].

We found a surprisingly low amount of synteny between *P. placenta* and the other two basidiomycetes. Only 41.3% and 33.8% of *P. placenta* genes were syntenic to those of *P. chrysosporium* and *L. bicolor*, respectively (**Table A2.1)**. Also, for comparison, we calculated the number of syntenic genes to be 44.7% between *L. bicolor* and *P. chrysosporium.* This indicated that there has been a high level of gene movement since the last common ancestor of the three basidiomycetes; however, it is unclear what would be expected. For

comparison, the amount of synteny found here was similar to a comparison between *Trichoderma reesei* and *Aspergillus nidulans*, the former being from the order Hypocreales and the latter a member of the order Eurotiomycetes (data not shown).

An intriguing finding was the high degree of overlap between the regions that were syntenic to both *L. bicolor* and *P. chrysosporium* (**Table A2.2**). This finding suggested those regions were resistant to change. This overlap of 14.3 total Mb and 3,575 *P. placenta* genes included 65.5% of the syntenic regions between *P. placenta* and *P. chrysosporium* and 81.5% of the syntenic regions between *P. placenta* and *L. bicolor*. This analysis also identified regions in the *P. placenta* genome that might be prone to reorganization.

To determine if genes with a particular function may have played a role in the restructuring of the genome, we annotated all predicted proteins in the three genomes with the Interpro[8]□ database to search for gene products with Interpro domains that were within or outside syntenic regions in the *P. placenta* genome. In **Table A2.3,** we present the results of the largest differences between Interpro families in and out of synteny with both *P. chrysosoporium* and *L. bicolor*. Similar to the analysis of Interpro domains and relationships to other genomes (see Supplementary Section "Addendum 1:Orthologs and Ka/Ks analysis of alleles"), we found that zinc fingers and other proteins with DNA binding domains were in the regions of no synteny, as were the cytochrome P450 categories. It is important to note that large regions of cytochrome P450 gene clusters have been discovered in the *P. placenta* genome (see Chapter 2), similar to those found in *P. chrysosporium* [5]. However, it is apparent that the same genes have not remained in clusters, or we would have found those genes in syntenic regions in this analysis. This was similar to findings in the *Trichoderma reesei* genome[4], where clusters of genes involved in biomass decomposition

were found to cluster outside of syntenic areas, suggesting that gene movement may have been favored by selective pressures.

Analyzing genes with respect to syntenic regions for trends in allele differences might provide other clues to why genome structures change. For this purpose, we again used the Ka, Ks and Ka/Ks scores (as described in Supplementary Note "Chapter 2, Addendum: Orthologs and Ka/Ks analysis of alleles") to assess the difference in alleles that lie within the two regions (overall results are in **Table 2.4**). Similar to what we showed in the "Chapter 2, Addendum: Orthologs and Ka/Ks analysis of alleles" section, we found that there were surprisingly more genes represented by allele pairs in the syntenic regions than the gap regions (see **Table A2.5**). For further analysis, we assumed that genes without preserved allele pairs in the assembly were alleles with 100% identity to each other, and thus the assembly algorithm was unable to separate them (the same correction as in Supplementary Section "Chapter 1, Addendum: Orthologs and Ka/Ks analysis of alleles"). When we analyzed alleles for Ka/Ks differences for genes that were within and outside syntenic regions, we found statistical significance (p-value less than .05) in both Ka (non-synonymous substitution) and Ka/Ks (non-synonymous divided by synonymous substitution). There were no significant differences in the comparison of *P. placenta* to *P. chrysosporium* gaps and syntenic regions. However, there were significant differences when we compared genes that were in gaps or syntenic regions between *P. placenta* and *L. bicolor.* This suggests that large regions of the genome are undergoing similar changes in nucleotide substitution rates. Further population level studies are needed for this group of fungi to determine how alleles are changing at this level.

References

1.   Seoighe C, Federspiel N, Jones T, Hansen N, Bivolarovic V, Surzycki R, Tamse R, Komp C, Hulzar L, Davis R, Scherer S, Tait E, Shaw D, Harris D, Murphy L, Oliver K, Taylor K, Rajandream M, Barrell B, Wolfe K: **Prevalence of small inversions in yeast gene order evolution**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**:14433-14437.

2.   Eichler EE, Sankoff D: **Structural Dynamics of Eukaryotic Chromosome Evolution**. *Science* 2003, **301**:793-797.

3.   Zhu H, Blackmon BP, Sasinowski M, Dean RA: **Physical Map and Organization of Chromosome 7 in the Rice Blast Fungus,** *Magnaporthe grisea*. *Genome Res.* 1999, **9**:739-750.

4.   Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, Chapman J, Chertkov O, Coutinho PM, Cullen D, Danchin EGJ, Grigoriev IV, Harris P, Jackson M, Kubicek CP, Han CS, Ho I, Larrondo LF, de Leon AL, Magnuson JK, Merino S, Misra M, Nelson B, Putnam N, Robbertse B, Salamov AA, Schmoll M, Terry A, Thayer N, Westerholm-Parvinen A, Schoch CL, Yao J, Barbote R, Nelson MA, Detter C, Bruce D, Kuske CR, Xie G, Richardson P, Rokhsar DS, Lucas SM, Rubin EM, Dunn-Coleman N, Ward M, Brettin TS: **Genome sequencing and analysis of the biomass-degrading fungus** *Trichoderma reesei* **(syn.** *Hypocrea jecorina***)**. *Nat Biotech* 2008, **26**:553-560.

5.   Martinez D, Larrondo L, Putnam N, Gelpke M, Huang K, Chapman J, Helfenbein K, Ramaiya P, Detter J, Larimer F, Coutinho P, Henrissat B, Berka R, Cullen D, Rokhsar D: **Genome sequence of the lignocellulose degrading fungus** *Phanerochaete chrysosporium* **strain RP78**. *Nature Biotechnology* 2004, **22**:695-700.

6.   Martin F, Aerts A, Ahren D, Brun A, Danchin EGJ, Duchaussoy F, Gibon J, Kohler A, Lindquist E, Pereda V, Salamov A, Shapiro HJ, Wuyts J, Blaudez D, Buee M, Brokstein P, Canback B, Cohen D, Courty PE, Coutinho PM, Delaruelle C, Detter JC, Deveau A, DiFazio S, Duplessis S, Fraissinet-Tachet L, Lucic E, Frey-Klett P, Fourrey C, Feussner I, Gay G, Grimwood J, Hoegger PJ, Jain P, Kilaru S, Labbe J, Lin YC, Legue V, Le Tacon F, Marmeisse R, Melayah D, Montanini B, Muratet M, Nehls U, Niculita-Hirzel H, Secq MPO, Peter M, Quesneville H, Rajashekar B, Reich M, Rouhier N, Schmutz J, Yin T, Chalot M, Henrissat B, Kues U, Lucas S, Van de Peer Y, Podila GK, Polle A, Pukkila PJ, Richardson PM, Rouze P, Sanders IR, Stajich JE, Tunlid A, Tuskan G, Grigoriev IV: **The genome of** *Laccaria bicolor* **provides insights into mycorrhizal symbiosis**. *Nature* 2008, **452**:88-92.

7.   **The Statistical Analysis of Spatially Clustered Genes under the Maximum Gap Criterion** [http://www.liebertonline.com/doi/abs/10.1089/cmb.2005.12.1083].

8.   Zdobnov E, Apweiler R: **InterProScan - an integration platform for the signature-recognition methods in InterPro**. *Bioinformatics* 2001, **17**:847-848.

Table A2.1. Features of syntenic regions calculated in the study. The last column represents the synteny of Laccaria bicolor to P. chrysosporium, as the percentages would be much lower if the opposite mapping was performed, due to the unusual nature of the much larger L. bicolor genome[6].

metric	P. placenta vs P. chrysosporium	P. placenta vs. L. bicolor	P. chrysosporium vs. L. bicolor
Number / percent of genes	5045 / 41.3%	4137 / 33.8%	4487 / 44.7%
Number / percent of orthologs	3398 / 36.2%	2844 / 29.8%	3029 / 30.9%
Percent of genome covered	33.6%	27%	39.9%
Number of syntenic blocks	391	373	333

Table A2.2. Overlap in the syntenic regions of *P. placenta* as compared to *P. chrysosporium* and *L. bicolor*.

Conserved synteny of both P. chrysosprium & L. bicolor to P. placenta	length in nucleotides	%
Length of shared synteny / percent of *P. placenta* genome	14257290	22.0%
L. bicolor total syntenic length	17486682	81.5%
P. chrysosporium total syntenic length	21762190	65.5%
conserved number of genes / percent of *P. placenta* genes	3575	29.2%

Table A2.3. Differences in the number of genes with domains identified by Interpro. *Postia placenta* genes in syntenic regions versus genes in gaps compared to both *P. chrysosporium* and *L. bicolor* are presented. The second and third columns indicate the proportion of all genes annotated in the genome.

P. placenta vs. P. chrysosporium			
Unique to Syntenic regions	**Prop. of annot. Genes**	**Greater in Syntenic regions**	**Prop. of annot. Genes**
IPR001683::Phox-like	0.0013872	IPR001680::G-protein beta WD-40 repeat	0.0079544
IPR000426::Proteasome alpha-subunit,IPR001353::20S proteasome, A and B subunits	0.0013872	IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase,IPR001245::Tyrosine protein kinase,IPR008271::Serine/threonine protein kinase, active site	0.0068832
IPR002041::GTP-binding nuclear protein Ran,IPR005225::Small GTP-binding protein domain,IPR001806::Ras GTPase,IPR003579::Ras small GTPase, Rab type,IPR003577::Ras small GTPase, Ras type,IPR003578::Ras small GTPase, Rho type	0.0013872	IPR000608::Ubiquitin-conjugating enzymes	0.0033476
IPR004274::NLI interacting factor	0.0009909	IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	0.0033374
IPR003593::AAA ATPase,IPR003959::AAA ATPase, central region,IPR003960::AAA-protein subdomain,IPR005937::26S proteasome subunit P45	0.0009909	IPR011545::DEAD/DEAH box helicase, N-terminal,IPR000629::ATP-dependent helicase, DEAD-box,IPR001410::DEAD/DEAH box helicase,IPR001650::Helicase, C-terminal	0.0026352
IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase,IPR001245::Tyrosine protein kinase,IPR000961::Protein kinase, C-terminal,IPR008271::Serine/threonine protein kinase, active site	0.0009909	IPR001993::Mitochondrial substrate carrier	0.0021586
IPR004087::KH,IPR004088::KH, type 1	0.0009909	IPR001138::Fungal transcriptional regulatory protein, N-terminal,IPR007219::Fungal specific transcription factor	0.0018801
IPR002068::Heat shock protein Hsp20	0.0009909	IPR002893::Zn-finger, MYND type	0.0015051
IPR003000::Silent information regulator protein Sir2	0.0009909	IPR000717::Proteasome component region PCI	0.0015051
IPR008011::Complex 1 LYR protein	0.0007927	IPR001163::Small nuclear ribonucleoprotein (Sm protein),IPR006649::snRNP	0.0014248
IPR000322::Glycoside hydrolase, family 31	0.0007927	IPR005225::Small GTP-binding protein domain,IPR001806::Ras GTPase,IPR003579::Ras small GTPase, Rab type,IPR003577::Ras small GTPase, Ras type,IPR003578::Ras small GTPase, Rho type	0.0014248
IPR004345::TB2/DP1 and HVA22 related protein	0.0007927	IPR000051::SAM (and some other nucleotide) binding motif	0.0014035
IPR000555::Mov34/MPN/PAD-1,IPR003639::Mov34-1	0.0007927	IPR005024::Snf7	0.0012480
IPR003397::Mitochondrial import inner membrane translocase, subunit Tim17/22	0.0007927	IPR000834::Peptidase M14, carboxypeptidase A	0.0012480
IPR001930::Peptidase M1, membrane alanine aminopeptidase,IPR006025::Peptidase M, neutral zinc metallopeptidases, zinc-binding site	0.0007927	IPR001841::Zn-finger, RING	0.0012429
IPR003527::MAP kinase,IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase,IPR001245::Tyrosine protein kinase,IPR008271::Serine/threonine protein kinase, active site,IPR008352::p38 MAP kinase	0.0007927	IPR002130::Peptidyl-prolyl cis-trans isomerase, cyclophilin type	0.0012266
IPR002553::Adaptin, N-terminal	0.0007927	IPR005829::Sugar transporter superfamily,IPR007114::Major facilitator superfamily,IPR003663::Sugar transporter,IPR005828::General substrate transporter	0.0012053
IPR001251::Cellular retinaldehyde-binding)/triple function, C-terminal,IPR008273::Cellular retinaldehyde-binding/triple function, N-terminal	0.0007927	IPR001005::Myb, DNA-binding	0.0011463
IPR001023::Heat shock protein Hsp70	0.0007927	IPR000195::RabGAP/TBC	0.0011088
IPR011550::Amidohydrolase-like	0.0007927	IPR001440::TPR repeat	0.0011036

P. placenta vs. P. chrysosporium			
Unique to Gap Regions	**Prop. of annot. Genes**	**Greater in Gap Regions**	**Prop. of annot. Genes**
IPR007087::Zn-finger, C2H2 type,IPR000345::Cytochrome c heme-binding site	0.0012531	IPR007708::Lariat debranching enzyme, C-terminal,IPR004843::Metallophosphoesterase	0.0001392
IPR002401::E-class P450, group I	0.0012531	IPR001878::Zn-finger, CCHC type	0.0361905
IPR003866::Isoflavone reductase	0.0009747	IPR007087::Zn-finger, C2H2 type	0.0130610
IPR003812::Filamentation induced by cAMP protein Fic	0.0008354	IPR001128::Cytochrome P450,IPR002401::E-class P450, group I	0.0109186
IPR011120::Neutral trehalase Ca2+ binding	0.0008354	IPR008266::Tyrosine protein kinase, active site	0.0051945
IPR002197::Helix-turn-helix, Fis-type	0.0006962	IPR001810::Cyclin-like F-box	0.0038226
IPR002889::Carbohydrate-binding WSC	0.0006962	IPR000719::Protein kinase,IPR002290::Serine/ threonine protein kinase	0.0033630
IPR001466::Beta-lactamase	0.0006962	IPR001395::Aldo/keto reductase	0.0023611
IPR003333::Cyclopropane-fatty-acyl-phospholipid synthase,IPR000051::SAM (and some other nucleotide) binding motif	0.0006962	IPR000719::Protein kinase	0.0021313
IPR007568::RTA1 like protein	0.0006962	IPR001248::Permease for cytosine/purines, uracil, thiamine, allantoin	0.0016546
IPR003042::Aromatic-ring hydroxylase,IPR002938::Monooxygenase, FAD-binding,IPR000733::Flavoprotein monooxygenase	0.0006962	IPR000172::Glucose-methanol-choline oxidoreductase,IPR007867::GMC oxidoreductase	0.0014564
IPR000585::Hemopexin	0.0005569	IPR006209::EGF-like	0.0014137
IPR000432::DNA mismatch repair protein MutS, C-terminal	0.0005569	IPR007269::Isoprenylcysteine carboxyl methyltransferase	0.0012745
IPR001005::Myb, DNA-binding,IPR001155::NADH:flavin oxidoreductase/NADH oxidase	0.0005569	IPR002086::Aldehyde dehydrogenase	0.0011190
IPR000379::Esterase/lipase/thioesterase,IPR002410::Peptidase S33, prolyl aminopeptidase,IPR005945::Peptidase S33, tricorn interacting factor 1,IPR000073::Alpha/beta hydrolase fold,IPR003089::Alpha/beta hydrolase0.000556947925368978		IPR001138::Fungal transcriptional regulatory protein, N-terminal	0.0010387
IPR001128::Cytochrome P450,IPR002401::E-class P450, group I,IPR008263::Glycoside hydrolase, family 16, active site	0.0005569	IPR001155::NADH:flavin oxidoreductase/NADH oxidase	0.0010174
IPR002016::Haem peroxidase, plant/fungal/bacterial	0.0004177	IPR001128::Cytochrome P450,IPR002403::E-class P450, group IV	0.0008995
IPR011118::Tannase and feruloyl esterase	0.0004177	IPR002110::Ankyrin	0.0008619
IPR001901::Protein secE/sec61-gamma protein	0.0004177	IPR001461::Peptidase A1, pepsin	0.0008568
IPR005197::Glycoside hydrolase, family 71	0.0004177	IPR000210::BTB/POZ	0.0008457

P. placenta vs L. bicolor			
Unique to syntenic regions	**Prop. of annot. Genes**	**Greater in syntenic regions**	**Prop. of annot. Genes**
IPR001410::DEAD/DEAH box helicase,IPR001650::Helicase, C-terminal,IPR000330::SNF2-related	0.0019338	IPR001680::G-protein beta WD-40 repeat	0.0071587
IPR004274::NLI interacting factor	0.0012086	IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase,IPR001245::Tyrosine protein kinase,IPR008271::Serine/threonine protein kinase, active site	0.0052140
IPR003593::AAA ATPase,IPR003959::AAA ATPase, central region,IPR003960::AAA-protein subdomain,IPR005937::26S proteasome subunit P45	0.0012086	IPR000504::RNA-binding region RNP-1 (RNA recognition motif)	0.0051341
IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase,IPR001245::Tyrosine protein kinase,IPR000961::Protein kinase, C-terminal,IPR008271::Serine/threonine protein kinase, active site	0.0012086	IPR000608::Ubiquitin-conjugating enzymes	0.0038566
IPR008011::Complex 1 LYR protein	0.0009669	IPR011545::DEAD/DEAH box helicase, N-terminal,IPR000629::ATP-dependent helicase, DEAD-box,IPR001410::DEAD/DEAH box helicase,IPR001650::Helicase, C-terminal	0.0032605
IPR003527::MAP kinase,IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase,IPR001245::Tyrosine protein kinase,IPR008271::Serine/threonine protein kinase, active site,IPR008352::p38 MAP kinase	0.0009669	IPR001440::TPR repeat	0.0026316
IPR001494::Importin-beta, N-terminal	0.0009669	IPR001993::Mitochondrial substrate carrier	0.0024063
IPR002041::GTP-binding nuclear protein Ran,IPR005225::Small GTP-binding protein domain,IPR001806::Ras GTPase,IPR003579::Ras small GTPase, Rab type,IPR003577::Ras small GTPase, Ras type,IPR003578::Ras small GTPase, Rho type,IPR002078::Sigma-54 factor, interaction region	0.0009669	IPR001841::Zn-finger, RING	0.0020246
IPR001023::Heat shock protein Hsp70	0.0009669	IPR001163::Small nuclear ribonucleoprotein (Sm protein),IPR006649::snRNP	0.0019228
IPR003397::Mitochondrial import inner membrane translocase, subunit Tim17/22	0.0009669	IPR001005::Myb, DNA-binding	0.0016756
IPR003892::Ubiquitin system component Cue	0.0009669	IPR000051::SAM (and some other nucleotide) binding motif	0.0016702
IPR006688::ADP-ribosylation factor,IPR005225::Small GTP-binding protein domain,IPR001806::Ras GTPase,IPR006687::GTP-binding protein SAR1,IPR006689::ARF/SAR superfamily	0.0009669	IPR004827::Basic-leucine zipper (bZIP) transcription factor	0.0015685
IPR003511::DNA-binding HORMA	0.0007252	IPR005024::Snf7	0.0015685
IPR010400::Protein of unknown function DUF1000	0.0007252	IPR001138::Fungal transcriptional regulatory protein, N-terminal,IPR007219::Fungal specific transcription factor	0.0014285
IPR001451::Bacterial transferase hexapeptide repeat	0.0007252	IPR002041::GTP-binding nuclear protein Ran,IPR005225::Small GTP-binding protein domain,IPR001806::Ras GTPase,IPR003579::Ras small GTPase, Rab type,IPR003577::Ras small GTPase, Ras type,IPR003578::Ras small GTPase, Rho type	0.0013267
IPR002226::Catalase	0.0007252	IPR001394::Peptidase C19, ubiquitin carboxyl-terminal hydrolase 2	0.0013267
IPR001296::Glycosyl transferase, group 1	0.0007252	IPR002423::Chaperonin Cpn60/TCP-1,IPR002194::Chaperonin TCP-1,IPR001844::Chaperonin Cpn60	0.0013267
IPR007124::Histone-fold/TFIID-TAF/NF-Y,IPR000558::Histone H2B,IPR007125::Histone core	0.0007252	IPR000426::Proteasome alpha-subunit,IPR001353::20S proteasome, A and B subunits	0.0013267
IPR001965::Zn-finger-like, PHD finger	0.0007252	IPR002130::Peptidyl-prolyl cis-trans isomerase, cyclophilin type	0.0013158
IPR003084::Histone deacetylase,IPR000286::Histone deacetylase superfamily	0.0007252	IPR000717::Proteasome component region PCI	0.0011977

P. placenta vs L. bicolor			
Unique to gap regions	**Prop. of annot. Genes**	**Greater in gap regions**	**Prop. of annot. Genes**
IPR007269::Isoprenylcysteine carboxyl methyltransferase	0.0017303	IPR001878::Zn-finger, CCHC type	0.0317002
IPR005630::Terpene synthase, metal-binding	0.0011123	IPR007087::Zn-finger, C2H2 type	0.0130321
IPR007087::Zn-finger, C2H2 type,IPR000345::Cytochrome c heme-binding site	0.0011123	IPR001128::Cytochrome P450,IPR002401::E-class P450, group I	0.0116999
IPR002401::E-class P450, group I	0.0011123	IPR008266::Tyrosine protein kinase, active site	0.0047130
IPR000772::Ricin B lectin	0.0009888	IPR001810::Cyclin-like F-box	0.0045751
IPR000194::H+-transporting two-sector ATPase, alpha/beta subunit, central region	0.0008652	IPR002347::Glucose/ribitol dehydrogenase,IPR002198::Short-chain dehydrogenase/reductase SDR	0.0038172
IPR002916::Ferric reductase-like transmembrane component	0.0008652	IPR000719::Protein kinase,IPR002290::Serine/threonine protein kinase	0.0024883
IPR003866::Isoflavone reductase	0.0008652	IPR000210::BTB/POZ	0.0023866
IPR003812::Filamentation induced by cAMP protein Fic	0.0007416	IPR000719::Protein kinase	0.0023592
IPR002018::Carboxylesterase, type B	0.0007416	uracil, thiamine, allantoin	0.0022356
IPR011120::Neutral trehalase Ca2+ binding	0.0007416	oxidoreductase	0.0019939
IPR002197::Helix-turn-helix, Fis-type	0.0006180	IPR001395::Aldo/keto reductase	0.0019414
IPR002889::Carbohydrate-binding WSC	0.0006180	oxidoreductase/NADH oxidase	0.0017358
IPR001466::Beta-lactamase	0.0006180	IPR006209::EGF-like	0.0014886
IPR001736::Phospholipase D/Transphosphatidylase	0.0006180	IPR002086::Aldehyde dehydrogenase	0.0012633
IPR007114::Major facilitator superfamily,IPR001411::Tetracycline resistance protein TetB	0.0004944	IPR007114::Major facilitator superfamily	0.0009636
IPR000585::Hemopexin	0.0004944	IPR000910::HMG1/2 (high mobility group) box	0.0008925
IPR000432::DNA mismatch repair protein MutS, C-terminal	0.0004944	IPR001138::Fungal transcriptional regulatory protein, N-terminal	0.0008925
IPR001005::Myb, DNA-binding,IPR001155::NADH:flavin oxidoreductase/NADH oxidase	0.0004944	IPR000873::AMP-dependent synthetase and ligase	0.0008870
IPR004304::Acetamidase/Formamidase	0.0004944	IPR000719::Protein kinase,IPR008266::Tyrosine protein kinase, active site	0.0008706

Table A2.4. Table showing the number of genes and percentage of genome with an allele pair that lie in syntenic and non-syntenic (gap) regions. Note the increased representation in number of genes in the syntenic as opposed to the gap regions.

Genome compared to P. placenta	gap	synteny
P. chrysosporium (number of genes/percentage of genome)	2317 / 32.26%	2513 / 49.8%
L. bicolor (number of genes/percentage of genome)	2859 / 35.3%	1971 / 47.6%

Table A2.5. Comparison of Ka, Ks and Ka/Ks values of genes with alleles that fall into the syntenic versus gap regions. For definitions of various measures, see Supplementary Note "Homologs and Allele comparisons".

Ka/Ks					
overall data	Ka	Ks	Ka/Ks	protein identity	nucleotide identity
Syntenic with *P. chrysosporium*	0.0317	0.0832	0.7591	96.8%	97.2%
Gap with *P. chrysosporium*	0.0357	0.0784	0.4586	96.4%	97.1%
Syntenic with *L. bicolor*	0.0280	0.0749	0.7363	97.2%	97.4%
Gap with *L. bicolor*	0.0372	0.0832	0.5040	96.3%	97.0%

CHAPTER 3

ADAPTIVE GENE CLUSTERING IN THE FUNGAL KINGDOM

Abstract

Previous studies have suggested that the location of genes in genomes is not random; instead they may be organized in a way that is beneficial to cellular processes and the organism. While a few studies have investigated the organization of genes on a whole genome scale, they were limited in the functions of genes used in the search and in the number and type of genomes searched. With the recent explosion of available fungal genomes and tools to automatically annotate many genes in a short period of time, it is now possible to obtain a global view of the level of clustering in the genomes of an entire kingdom. To find gene clusters in many genomes, we have constructed a robust and flexible algorithm that runs in trivial time. In parallel, we have annotated 72 fungal genomes using four automated annotation tools that provide information about protein function, protein targeting, involvement in biochemical pathways and paralogous gene families. We used the clustering algorithm to search for clusters from the four annotation categories. We discovered that all the genomes contained clusters of related genes, and that in several cases the clusters included genes involved in processes that were specific to the species in which they are found. This has dramatically expanded our knowledge of both the types of clusters and the number of genomes known to contain clusters. This study has generated information that will assist researchers in addressing many questions central to molecular and cell biology as well as evolutionary studies. To this end, the thousands of clusters we have discovered are available for download at kiddomics.com/.

Introduction

In recent years, there has been a dramatic increase in the number of complete (or nearly complete) genome sequences available from many types of organisms [1]. Analysis of the stream of genomic data has altered our understanding of what happens to a genome over time. Before the genomic revolution, many molecular biologists gave maximum importance to nucleotide mutations when considering the types of changes important in evolution. Subsequent thorough analyses have implicated gene duplications [2], transposable elements [3, 4] segmental duplications [5] and gene translocations [6, 7] as additional powerful forces of genome evolution; some such forces might be so catastrophic as to cause speciation, independent of the environment [8].

Of these forces, the movement of genes is the least understood, in part due to the intractability of tracking gene movements. While confirmation of genomic changes came only after whole genome sequences were available, bold researchers proposed such events long before the genomic era, at a scale ranging from the single gene duplication level to whole genome duplication [9]. While some cause-and-effect relationships have been established between genome changes and phenotypes [10, 11], there has been little study as to the impact of gene function on genome structure, if any, and none outside of a handful of model organisms.

Have Clusters of Genes Previously Been Discovered in Genomes?

There are many well known cases of spatial gene clusters. In bacteria there is a mechanism to directly support clusters of related genes, i.e., operons that must be transcribed together. Examples include the well known *lac* (*lactose*) operon; also, cellulosome genes (encoding the proteins that form a pseudoorganelle of cellulase degrading proteins in bacteria) are

often found in one or more operons [12, 13]. The existence of gene clusters in eukaryotes appears to be unequivocal, as many genes of related function, biochemical pathway or coregulation are known (reviewed in Hurst et al. (2004) [6]). In fungi in particular, there have been years of research into secondary metabolism clusters, encoding gene products that work in concert to produce a metabolite, usually anti-microbial. Many secondary metabolites are important drugs, such as statins from *Aspergillus terreus* [14]. Fungi also contain clusters of genes that are similar in sequence (termed paralogs). These have been found in basidiomycetes that are studied for their unique ability to degrade lignin, a large and diverse aromatic organic compound. It appears to be difficult to "engineer" an enzyme to attack lignin, which must be decomposed by oxidative rather than hydrolytic means. Many of the paralogous monoxygenase genes are found clustered in *Phanerochaete chrysosporium* [15] and *Coprinopsis cinerea* (Jason Stajich, personal communication). Finally, co-expression clusters in the yeast *Saccharomyces cerevisiae* and mammals have been discovered through transcriptomic studies[16, 17].

To date there are only two major studies of gene clustering in eukaryotes, one limited to the model organism quintet (*Saccharomyces cerevisiae*, *Homo sapiens, Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Drosophila melanogaster*); in the more recent study*, Escherichia coli, Danio rerio* and *Mus musculus* were added) [18, 19]. While these are quite important organisms in the cell and molecular biology field, they represent a tiny fraction of the organisms on the planet. We must cast a wider net to understand the connections between gene function, gene history, gene clusters and evolution.

Lee and Sonnhammer (2003) [18] produced the first analysis of gene clustering in eukaryotes. As is sometimes the case in initial attempts, the method was somewhat

awkward and seemed to suffer from a lack of statistical significance. The analysis involved little more than calculating the distance between genes divided by the length of the chromosome. Expected and actual scores were calculated and compared for each gene family under investigation. Also, this study examined only the clustering of genes producing proteins involved in biochemical pathways, and not any of the other types of clusters that may be present in the genome. Still, some of the results from the study were intriguing; for example, an astonishing 98% of the biochemical pathways of the yeast appear to be present in clusters as defined by KEGG EC (Enzyme Commission) code maps [18, 20]. We included biochemical pathways in our study and will compare our data to the previous study.

The recent study of Yi et al. (2007) [19] employed a more robust method in the analysis, but the study used only Gene Ontology (GO) [21] results, which do not provide as much coverage as other methods of functional characterization such as Interpro [22]. In addition, the variability in the depth of the GO tree that is scanned gives mixed and overlapping clusters that have to be resolved after the initial clustering pass, and this taints the final results with uncertainty. Nevertheless, the hypergeometric model used by Yi and colleagues is appropriate for this type of data, and so was used in a similar manner in this study.

RESULTS

Design and Implementation of an Algorithm for the Discovery of Gene Clustering in Eukaryotes

The first step in a computational analysis project is to produce a robust and flexible piece of software. The positive aspects of the method in used by Yi and colleagues [19] are that it is fairly straightforward and that it uses the hypergeometric distribution as a test for

significance. The hypergeometric distribution as a test is also attractive, as one of the parameters is the size of the gene family, and the analysis accounts for different sizes of gene families or annotation categories. We developed a method similar to that of Yi et al. (2007) [19]. However, we reduced the complication of the Gene Ontology [21] graph traversal, and applied a simple "identically annotated genes are a cluster" strategy, by defining a cluster of genes as those from a particular annotation category (termed target genes) located within a certain number of genes (defined below) on a scaffold or piece of scaffold and bounded by genes from the same annotation category.

Since the hypergeometric distribution takes into account all the genes in the genome as background, we define the number of genes in the genome as **n**. A subset of **n**, the number of genes in the genome from a particular category that could cluster, is termed **n'** (defined as target genes). The total number of genes on the current scaffold is denoted as **c**. For each cluster on a scaffold or contig, n'_i is called k'_i or the genes in a cluster from annotation category **n',** and all non-target genes in a cluster are termed **k**. The algorithm begins as described by the pseudocode in **Box 1** by locating the first **n'** gene at the furthest left of each chromosome piece and adding genes into the cluster as the algorithm scans to the right. The starting point of a cluster is termed k'_1 and genes that are not from **n'** are added into the cluster as k_i, $k_i + 1$ ... k_c. If a gene from the same set of **n'** genes is reached, this is added as k'_{i+1} and the significance is checked using a Perl to R (cran.r-project.org) bridge. This is to utilize the primitives dealing with hypergeometric models in R. If the hypergeometric p-value is less than a significance cutoff alpha, the process is repeated and k_i genes are added into the cluster until the next n'_i genes are reached or until **c** number of genes are added. If instead the p-value is greater than the alpha value, the last **k'** gene is trimmed off and all the intervening k_i genes are removed, the algorithm returns the last **k'**

gene that formed the most significant cluster, and the last **k'** gene that created an insignificant cluster is turned into the **k'$_1$** (first gene of the potential new cluster). The algorithm proceeds through all chromosome pieces in the genome with at least two **n'** genes. Single genes are considered insignificant. The algorithm complexity is linear, as the theoretical maximum time if all genes in the genome were in clusters, would thus be O(n), adding the constant time to complete the Perl to R bridge for every 2 **n'** that begin and end a cluster. In practice on an AMD Phenom 9750 quad core CPU with eight gigabytes of RAM, no genome in our study took longer than 28 minutes to complete and ninety percent of the genomes took no longer than fifteen minutes to complete.

Pan Fungal Discovery of Gene Clusters

Determining which genes can form clusters is somewhat arbitrary, and thus our approach made use of several different annotation categories in a comprehensive search for gene clusters in filamentous fungi. This was an attempt to capture all possible ways in which related genes can be collocated in genomic space that may be favored through selection and thus fix in the population. While Yi et al. (2007) [19] used Gene Ontology for functional inference of a gene, and Lee and Sonnhammer (2003)[18] used the KEGG [20] maps to associate EC codes (biochemical enzyme commission designations) to biochemical pathways, we used different databases more appropriate for analysis in fungi. In addition, we searched for clusters of other types of annotated genes, such as paralogous genes.

Genes with Similar Interpro Domains

While previous methods have utilized GO [21] and KEGG [20], we chose to annotate all genes from each genome in **Table 1** with Interpro [22], as Interpro will locate functional matches for a higher portion of the genes in a genome. Also, we used the simple rule that

"genes with same annotation" are potential clusters. As there are shallow parent-child relationships in Interpro, we present data for both types of clusters that exist in the genomes. There is also the possibility that multiple domains can be returned for any single gene model. For the purpose of simplicity we consider each domain clusterable. Details on Interpro annotation are in the **Materials and Methods** section.

We applied the cluster finding program to all the Interpro annotated genomes listed in **Table 1**. We first analyzed clusters of genes whose protein products contained the same parent term. The algorithm located clusters in all genomes scanned, as shown in **Figure 1**. For comparison we calculated the proportion of genes annotated in the genome, proportion of genes annotated that were found to be clustered and proportion of genes in the genome found in clusters (**Figure 1**). The mean proportion of genes annotated across all genomes was fairly high (.64) and had the highest variation of the three measures (standard deviation = .123). The mean proportion of genes clustered that were annotated (genes that contain at least one Interpro assignment) in all genomes was .256, resulting in a mean proportion of .163 for all genes in all genomes. The variability in the proportion of annotated genes clustered and all genes clustered was far lower than the variability seen in the mean proportion of genes annotated (standard deviation = .050 and .038, respectively). This indicates that while sizable portions of the genomes were annotated, relatively few of the annotated genes could be found in clusters.

Similar to the Yi et al. study [19], this algorithm produced clusters that were fairly diffuse, with a high number of intervening genes dispersing the target genes. In **Supplementary Table 2** we show the number of clusters for each genome, as well as several other key metrics relevant to the overall performance of the algorithm. In all genomes, the mean

107

number of target genes per cluster was rather low, resulting in the overall average of 2.55 genes. We chose to include pairs, as in smaller genomes like those of fungi, many important gene families may only appear as pairs or triplets. The mean density score (target genes/non-target genes in cluster) was quite good; the mean for all genomes was .23, ranging from a low of .14 (*Magnapothe grisea* [23]) to a high of .43 (*Candida parapsilosis*). At first glance, this might be thought to reflect the rather large difference in genome size between filamentous fungi and the yeasts, as the former generally contain more than 11,000 genes while the latter comprises typically less than 6,000. However, the fungus with the next highest mean density of .40, *Coccidioides immitis* strain rmscc37031, has 10,043 genes. Thus mean density score does not appear to correlate well with genome size or gene number.

An interesting pattern emerges when we explore the upper and lower ranges in the metric "proportion of genes in the genome clustered" in the genomes shown in **Figure 1**. The genome with the highest proportion was the basidiomycete *Phanerochaete chrysosporium* (.232) [15]. This is not surprising, as this genome was previously shown to contain large clusters of genes that are involved in lignin degradation (which were indeed recovered in our study as Interpro domain ID IPR001128:Cytochrome P450, totaling 47 genes). The ability to degrade lignin is a unique capability of whiterot fungi, for which *P. chrysosporium* is the model organism. On the other end of the spectrum, a scant .07 of genes in the genome of the basidiomycete and wheat pathogen *Puccinia graminis* were found to be in clusters. It is surprising to find that both ends of the spectrum were capped by basidiomycetes. Observing the top five fungi in this spectrum, we found that the next four genomes with the highest proportion of genes in clusters (*Neosartorya fischeri* (.216), *Aspergillus fumigatus* (.215), *Aspergillus oryzae* (.212) and *Aspergillus flavus* (.210)) are soil-borne fungi that are involved

largely in biomass degradation. While *N. fischeri* and *A. fumigatus* are opportunistic pathogens, they are usually found in soil and appear to be common food decay fungi [24]. On the opposite end of the spectrum, we found that the next four fungi with lowest proportions of clustered genes (*Batrachochytrium dendrobatidis* (.073), *Candida albicans* (.076), *Botrytis cinerea* (.078) and *Ustilago maydis* (.0878)) are all well known pathogens of animals and plants, as is *P. graminis.* It is also interesting to note that while the top of the spectrum contains many phylogenetically related fungi from the genus *Aspergillus*, the fungi with fewest clustered genes are not closely related at all. This may reflect the pressures of generalists that predominantly live in soil versus specialists that are the pathogens.

Clusters of Genes with Same Child Interpro Domains

The genomes in **Table 1** were annotated as described above, with the exception that we left the annotation as the child domain. The same values are shown: proportions of genome annotated, genes annotated that are clustered and genes in genome that are clustered. Again, all genomes contained some genes that can be found in clusters. The proportions of genes in the genome that were annotated and clustered are shown in **Figure 2**.

As might be expected from the shallow parent-child relationship in Interpro, the mean proportion (across all genomes) of genes in the genome clustered with child Interpro annotations (.17 with a mean standard deviation of .04) was nearly identical to those of the parent mapped clusters. The five genomes with the highest proportions of genes in the genome found in clusters and the bottom five were also nearly identical to that described above, with the only difference being the transposition of *Aspergillus fumigatus* and *Neosartorya fischeri.* This suggested that the work performed to map the child term back to the parent term may not have altered the results in any meaningful way. We will, however,

still refer to both sets of results in the remainder of our study as we feel it highlights different effects. We also note that the mean and standard deviation were slightly higher, which may be due to the change in the number of target genes per gene family, or **n'** (**k'** in clusters). This relationship is illustrated in **Supplementary Table 2**, as the mean difference between the parent Interpro mean and the child mean for each genome was .007, while the differences in mean densities was $9x10^{-4}$. It is worth mentioning that the sorting order was nearly identical for these relationships. Additionally, the number of clusters was always predictably larger for the child term data. Since the proportion and rank changed little, it appeared that the breaking of the parent terms down into the child simply had the effect of breaking the larger parent clusters into smaller clusters.

Clusters of Genes with the Same Localization Signals

An important group of genes are those targeted to the same cellular location. Most importantly in fungi, genes whose products are secreted play an important role and were previously shown to cluster in at least one fungus [25]. To identify the potential localization of these genes in the genome, we used WoLF-PSORT [26]. While WoLF-PSORT proposes a location for every protein in the genome, we instead decided to trim the number of genes annotated by comparing the score to that of another localization prediction method, Phobius [27]. This procedure is described in greater detail in Materials and Methods. The general effect, the drastic trimming of the number of annotated genes per genome, is visible as the red bars in **Figure 3**.

The proportion of genes in the genome that were clustered and annotated by WoLF-PSORT (**Figure 3**) was very similar to the proportion of genes that were clustered with similar Interpro domains (**Figures 1** and **2**). The mean difference in the proportion of genes that

were clustered in the genome was only .10, and the mean difference in genes annotated that were clustered was an even lower .01. Statistical tests for similarity showed that for 18 genomes the proportions are close enough to be considered the same. Still, for the majority of our data set the proportions were statistically different, although for those 18 genomes there is the possibility localization and function are related. This is somewhat surprising as the annotated proportion of the genome was far smaller for the WoLF-PSORT clusters (mean of .25) than for the Interpro clusters (mean of .64). The confounding factor may be that the number of target genes is much larger per category, as there only seven (cytoskeleton, cytosol, extracellular, mitochondrial, nuclear, peroxisomal and plasma membrane) potential locations to which a gene can be targeted, while there are more than three thousand Interpro terms with which a gene can be associated.

The combination of fewer annotated genes and similar percent of genome clustered had the expected effect of creating clusters that were on average more dense than those of the smaller Interpro domains. The mean density for the WoLF-PSORT clusters was .70, which was far greater than those of either the child or parent Interpro annotated genomes (both .23). Given the high density, it is tempting to conclude that genes targeted to similar locations are subject to higher pressure to be collocated in the genome than are genes that have the same or related functions; although more study is needed to evaluate this possibility.

Search for Clusters of Genes Involved in Biochemical Pathways

Lee and Sonnhammer [18] scanned several genomes (the traditional Model Organisms) for clusters of biochemical pathway genes. While their method has limitations as far as significance of the clusters, biochemical pathways are still important to search for potential

clustering. We used BioCyc [28] in place of the KEGG [20] pathway maps used in the Lee and Sonnhammer study to provide higher coverage and a rich detailed interface, as well as a built in annotation method to search for gene products involved in known biological pathways, as described in Materials and Methods.

With this annotation method we chose to search for clusters in two ways. The first search method was by individual pathways as identified by the Pathologic tool in BioCyc [28]. The second approach allowed for any gene producing a protein identified as involved in a biochemical pathway to be in a cluster with any other gene product involved in any biochemical pathway (simply put, pathway dependent and pathway independent, respectively). The former gave a glimpse of pressures for particular pathways to be in a cluster, while the latter allowed for the possibility that all biochemical pathways are actively under some sort of related pressure.

The pathway dependent results are shown in **Figure 4a**. One notices immediately that the proportion of genes that were annotated in the genome (mean = .06) was far lower than for the Interpro data (mean = .64) or the paralogous gene data (mean = .37) that are described below. Nevertheless, the proportion of annotated genes that were clustered was quite high (mean = .28). In fact, this proportion was higher than either the Interpro measure (parent mean = .25, child mean = .26) or that of paralogous gene clusters (mean = .18).

Several genomes had an unusually high number of genes whose products were in pathways (pathway dependent) that were clustered. Oddly, four of the top five genomes with the highest number of biochemical pathway genes clustered were of the genus *Coccidioides* (ref, sharpton et al), which represent four of the six *Coccidioides* genera included in our

study. The fungus with the highest proportion of genes clustered (.50) was *Coccidioides immitis* strain rs, followed by *Coccidioides posadasii* strain rmscc3488, *Coccidiodes immitis* strain rmscc23941 and then *Coccidioides posadasii* strain Silveria. While the two strains of *C. immitis* might be expected to have very similar clustered gene contents, it remains an open question as to why the other two *C. immitis* strains (h53841 at 21st and rmscc37031 at 42nd) ranked far lower.

It is tempting to draw a connection between the clustering of primary metabolism protein-producing genes and pathogenesis in the genus *Coccidioides*. The largest cluster in the *Coccidioides immitis* strain rs genome was an overlapping cluster of 3 pathways (PWY-561:superpathway of glyoxylate cycle, PWY-4302:aerobic respiration -- electron donor III, and PWY-3781:aerobic respiration -- electron donor II), which shared some of the same enzymes between them. In particular, that the PWY-4302:aerobic respiration -- electron donor III pathway utilizes a ferroprotein as the terminal protein (as opposed to cytochrome) is of interest with respect to pathogenesis. Recently Sharpton et al. (2009) [29] suggested, based on the analysis of several *Coccidioides* genomes, that these species have "acquired genes that enable the fungus to grow on metabolites found within a host, such as iron and nitrogen [29]." This is intriguing, as it suggests that the scavenging and use of iron may be involved in pathogenesis, and thus the fungus may be under pressure to keep the genes in a section of the genome that would allow rapid access to transcriptional machinery. However, the recently diverged non-pathogenic *Uncinocarpus reesii* [29] ranked 6th in the list, above two pathogenic *Coccidioides* strains, and contained the same clusters, albeit not to the same degree.

Surprisingly, when we removed the restriction that genes must cluster by annotated pathway, the percentage of annotated genes did not change dramatically (mean difference = .01, paired t-test p-value = .4). In 34 of the 72 genomes (.47), the proportion of annotated genes increased in the pathway independent experiment (**Figure 4b**). We had expected that the proportion should increase with fewer restrictions, however it appeared that this had little or no effect overall. We also did not expect that the proportion would drop for the independent results. This could be caused by the increase in the size of the gene family, which adjusts the probability of success in the hypergeometric distribution, our statistical model. This change could have made it more probable that a gene in a cluster could be found by chance, and might have resulted in the small change observed. Still, this surprising result suggested that if there is any pressure for clustering, it is at the level of the individual pathway and not simply any biochemical gene for any pathway.

In a previous study, Lee and Sonnhammer [18] reported that 98% of biochemical pathways showed significant clustering in the genome of *Saccharomyces cerevisiae*. In our study, the pathway dependent and independent analyses showed that the number of genes from biochemical pathways clustered in the *S. cerevisiae* genome were 21% and 20%, respectively. When the number of pathways (a measure that was comparable to that of the Lee and Sonnhammer study) with at least a pair of genes clustered was calculated, we found a similar percentage of 21%. We included two *S. cerevisiae* strains in our study, ATCC-AB972 that was the first eukaryotic genome sequenced [30] and the rm111a1 strain. In the rm111a1 strain, we found that the proportion of biochemical genes clustered was .21 in the pathway dependent experiment and a surprisingly lower .17 in the pathway independent analysis. Additionally, in the rm111a1 strain the percentage of biochemical

pathways was 25%. This was far lower in either strain than that found in the previous study by Lee and Sonnhammer.

There could be two reasons for the large difference in the percentage of biochemical pathways found to contain gene clusters in our work compared to the earlier study. First, the underlying model in the studies was very different. Whereas Lee and Sonnhammer used expected versus observed distributions when comparing a clustering "score" that was dependent on the length of the chromosome, we used the hypergeometric distribution to control for various problems such as change in number of genes in a pathway and differing number of genes in the genome. This appeared to have the effect that our study was more conservative and may have thrown out some results. The second reason could reflect the different number of pathways annotated between the two studies. In the Lee and Sonnhammer study, 105 pathways were used, whereas in our study we obtained 197 for the ATCC strain and 216 for the rm111a1 strain. This could be due to the source of information on biochemical pathways. The Lee and Sonnhammer study used the KEGG [20] pathways, while we used the Pathologic tool [28]. The two tools map out the reaction pathways in slightly different ways. KEGG clumps together some related pathways, while Pathologic tends to divide related pathways. We would expect to have overestimated the number of pathways clustered in our study, as it presented a smaller number of target genes to our hypergeometric model, which tends to deflate the p-value and would cause clusters to erroneously form. In any case, our smaller percentage of clustered pathways most likely represents a more stringent set of clustered pathways than that of the previous study.

Search for Clusters of Genes from Paralogous Gene Families

This data set was the result of finding paralogous genes with MCL [31] (Markov Clustering Algorithm). In short, MCL groups proteins together by sequence similarity and utilizes a graph approach to tease apart the groups into closely related subgraphs. The MCL grouping was performed as described in Arvas et al. (2007) [32]. Each group of genes that was found to be in the same paralogous group will be tested as to whether they spatially cluster in the genome. Unfortunately, due to time constraints and machine limitations, the data available for this study were not as extensive as for the other analyses, so only 44 genomes were available. However, the data were still valuable, and included many of the phylogenetically distinct species (but not as many of the different strains for some species). This was a nice measure of clustering as it addressed the question of paralogous genes maintaining their relative locations in the genome.

The overall structure of the related MCL clusters was similar to that of the Interpro results (**Figure 5**). This was not surprising, as it is likely that the majority of genes with the same Interpro annotation are related by high sequence similarity. Interpro, however, contains many genes with a domain in common that would be called different MCL groups or subgroups of the same MCL paralogous group. In the MCL gene cluster data, the mean number of target genes was 2.6 per cluster and the mean cluster density was .31, lower than the density of the WoLF-PSORT data. However, there was a higher variability than for the other analyses; the standard deviation was .12 for history cluster density, while for the other analyses the value ranged between .05 and .06. From this assessment, we can easily conclude that by and large most paralogous genes do not maintain their relative locations in the genome. However, this does not address the question of how old the genes are when

their location is changed. More analysis, outside the scope of this project, would be needed to address this question.

The genome with the highest mean density of clustering, at an astonishing .59 with respect to the MCL genome clusters, was *Botrytis cinerea*. This was intriguing, as *Botrytis* is a major pathogen of wine grapes and soft fruits such as strawberries [33]. It appeared that a high number of Cytochrome P450 domain containing genes were clustered in tandem duplicates, as well as a high number of peptidases with the same pairwise organization. Interestingly, these classes of genes were found to be upregulated during infection [34]. In fact, we found several genes annotated as aspartic peptidases that were also clustered as pairs in the genome (4 pairs). This added suspense, as these were found in a similar study   to encode a large percentage of the protein secreted by *B. cinerea* during infection. These two classes of proteins are suspected of enabling *B. cinerea* (a necrotroph) to evade host defenses and destabilize cell walls [35]. Monoxygenases (containing Cytochrome P450 domains) may also be involved in causing lesions of dead tissue that are usually caused by the plant as a defense mechanism (since some pathogens cannot live on the dead tissue and do not survive to continue infection) [36]. This draws an interesting parallel to the *Phanerochaete chrysosporium* use of similar proteins to bombard lignin in order to extract nutrients from wood. Similarly, another nectrotroph, *Sclerotinia sclerotiorum*, ranked near the top at 9th with .43 of the MCL annotated genes in clusters. There did not seem to be any preference for pathogenesis and clustering on this metric, as *Magnaporthe grisea* ranked near the bottom with .11 of the MCL genes clustered. Clearly further studies are needed to distinguish between these two possibilities.

DISCUSSION

In the previous section we discovered that all genomes contain clusters of related genes. We also showed that there is a high amount of variation in the proportion of clusters in different genomes. This raises two major questions: are the clusters heritable and how are different cluster categories distributed across the genomes? To answer these questions we used a method similar to the biclustering method, which clusters data on two axes at the same time, and reorders the columns based on the outcome. The method we chose was hierarchical clustering with complete linkage, as this is a preferred method for placing genomes (and target gene categories) that have the most similar profiles in content together. In order to create the data set, values had to be filled in for genomes that did not contain any genes from a particular family or target gene set. We tested whether it was best to code this case as minus one or zero. There were concerns that a zero could be confused with cases where genes were present in a genome for a particular targeting category, but no clustering of the genes was found. In these tests, we discovered that there was no difference between the codings of zero and minus one. Thus we used a zero to indicate both that there were no genes annotated for a particular category, and that genes were annotated for a particular category but no clustering was found (data not shown).

Interpro Clusters

We used the basic "heatmap" function in R (cran.r-project.org) to produce the heatmap in Figure 6a, which shows the effect of clustering both the Interpro categories (x-axis) and the genomes (y-axis). We then used the resulting hierarchical tree (Figure 6b) for comparison with the phylogenetic tree in the James et al. (2006) [37] study. It is important to note that the tree in Figure 6b is not a phylogenetic tree, but is a graphical representation of clustering the genomes with the most similar proportion of Interpro annotated genes next to each

other. Another point of difference between a hierarchical clustering tree and a phylogenetic tree is that the branch lengths do not signify anything, as they often do in phylogenetic trees.

While most of the genomes clustered according to their phylogenetic order (suggesting that the clusters are inherited), there were several genomes whose positions in the hierarchical tree were very different than expected. Notably, the *Coccidioides* genomes fell in a pattern similar to that shown in the results section above. The *C. immitis* strains h53841 and rmscc37031 fell into a different part of the tree than their respective relatives, more similar to a close relative in the Onygenales, *Paracoccidioides brasiliensis*. These three genomes fell into an interesting group, as it contained organisms that are pathogens of both plants and animals, and whose hierarchical clustering order was in drastic disagreement with the relationship expected by phylogenetic placement. This group included *Botrytis cinerea*, *Encephalitozoon cuniculi*, *Ustilago maydi*s and *Aspergillus nidulans*. This group was hierarchically clustered in both the parent and child data, as highlighted by the blue bracket in Figure 6b.

To find the largest differences between these genomes, we used the principle component analysis (PCA) method [38]. PCA is a matrix manipulation technique that provides information on which variables in a data set are contributing the most variability. The result is a set of principal components, the first of which contains the most variability, the second the next most, etc. We used the 12 genomes described above for this analysis, all *Coccidioides* genomes with the closely related *Uncinocarpus reesei.* We included *P. brasiliensis, B. cinerea, E. cuniculi, U. maydis and A. nidulans* to compare the differences between the expected and actual relationships in hierarchical clustering.

When we analyzed both the parent and child versions, PC1 (principle component one) contained 34% of the variability and PC2 13 %. We sorted through the top 25 loadings (which are Interpro categories) and found that the largest differences involved the group containing the *C. immitis* strains h53841 and rmscc37031, along with *B. cinerea, E. cuniculi, U. maydis and A. nidulans.* These genomes had far fewer clusters. This may represent a "long branch attractor" effect, where genomes that are simply very different from all the rest are placed together in the tree. However, the question still remains as to why these genomes, especially the *Coccidiodes* outliers and *Aspergillus nidulans*, had far lower levels of clustering.

To find the most dynamic categories in the Interpro clusters across all genomes, we again used the PCA technique and analyzed the top 25 loadings in PC1. To simplify the analysis, we collapsed the parent and child Interpro lists into one, as most of the terms appeared in both. This produced 35 unique Interpro terms. The top term, IPR013708 (and IPR006151 further down in the list), identified the shikimate/quinate pathway [39]. It has been known for over 30 years that the genes involved in the utilization of quinate appear in a tight cluster in the filamentous fungi. This served to validate our method. Also, this quinate utilization clustering appeared in most Eurotiomycetes and Pezizomycetes, but appeared to be missing in *Aspergillus nidulans* and some other fungal orders, including the Saccharomycotina.

Of the other categories in the Interpro data, we found that the highest number of domains in PC1 were likely involved in biomass degradation (12 Interpro categories), with secondary metabolism a close second (10 Interpro categories), as shown in Table 2. This is not surprising, as both types of proteins were found to cluster in the genome [11, 24, 25, 40].

Biomass degrading genes have a varied content across genomes, and might be tuned to the types of carbon sources to which the organisms have adapted. It is interesting that in the list of categories with the highest loadings that belong to biomass degrading enzymes, the pectin degrading Interpro domains were the most varied. It is thought that some pectin degrading enzymes might be involved in plant pathogenesis [41]. However, the heatmap in figure 6a does not suggest the genes encoding pectin degrading enzymes were only clustered in pathogens. The secondary metabolism genes were also a nice point of validation, as this group has been known to be clustered for many years [40].

Table 2 also revealed clusters of the HMG1 (high mobility group) gene products that are encoded by the mating-type loci of most filamentous fungi; these products control sexual types in the fungi. The variety in the organization and number of mating-type loci makes these genes one of the most variable groups in fungi [42].

On the opposite end of the spectrum were genes that were clustered across all genomes. To find these Interpro categories, we first found Interpro IDs that appeared in all genomes, and then sorted the categories from highest to lowest mean proportion of genes clustered. This identified only 23 parent Interpro terms that were found clustered across all genomes (Table 3). The top term, "IPR016196:Major facilitator superfamily general substrate transporter" typically identifies sugar transporters involved in sugar import into the cell for carbon sources and possibly for sensing [43]. This highlighted the importance of carbon metabolism in all fungi. The rest of the list contained many proteins that are essential for basic cellular metabolism, including domains that are involved in DNA/RNA interaction, signaling, and other transporter types. The DNA/RNA interaction domains appear to be involved in regulation (such as the zinc fingers, IPR015880 and IPR001841) and chromatin

maintenance (IPR001650 and IPR014021, helicases). The signaling molecules were identified by the numerous kinase interaction domains in Table 3.

Clusters of Genes Producing Proteins with Similar Targeting Signals Revealed Regions of Genomes with Concerted Function

We applied the clustering and PCA techniques described above to the genomes annotated with WoLF PSORT [26] for targeting locations. We applied the same filtering to the annotations as described in Materials and Methods. The heatmap (Figure 7a) and dendrogram (Figure 7b) show the effects of the hierarchical clustering of both the x and y axis, as for the Interpro analysis. In the Interpro experiment, there was general agreement between expected phylogenetic relationships and the hierarchical clustering results. With the WoLF PSORT clusters, however, this seemed to not be the case. The few groups that stayed together included the *C. immitis* strains h53841 and rmscc37031, which were again separate from the the rest of the genomes of the genus *Coccidioides,* which appeared in the dendrogram as two separate groups. Also, the two *Neurospora* genomes, *N. tetrasperma* and *N. crassa,* were grouped closely, as were most of the *Aspergillus* genomes.

The PCA analysis showed that PC1 contained 32% of the variation while PC2 contained 23%. The top loadings in PC1 were nuclear targeting followed by cytoskeleton. The bottom of the list included, perhaps surprisingly, plasma membrane and extracellular targeting. We had expected that there would be more generic cellular processes targeted to the cytoskeleton and nuclear compartment, while secreted proteins would be the most varied, as these might be dependent on carbon source utilization, pathogenesis and environmental sensing. However, this appeared to not be the case, as it was evident that all fungal

genomes contained a portion of proteins that were extracellularly targeted, while other cellular locations did not appear to be as consistent in their distribution of clustered genes.

To understand what is at the heart of the pattern behind the secretion targeting, we sampled the largest clusters in a random sample of genomes and collected the functions of the target genes in the clusters from Interpro. The largest cluster of genes that contained a secretion signal was a cluster of genes from *Fusarium graminearum* [44], which contained 196 genes encoding proteins containing a secretion signal. While this region spanned a considerable portion of the genome (1915 total genes, p-value 1.4e-12), it is worth describing here. This region, described in Supplementary Table 1, contained many glycoside hydrolase genes (28 unique) [45], eight unique pectin lyases (several with cellulose binding domains), nine unique oxidase genes, ten different esterases, six different peptidases and two different chitinases. Many of the genes listed were in multiple copies, and thus were partially responsible for the statistical clustering score. In addition to these genes which are obviously involved in plant biomass degradation, there were several genes encoding proteins that are uniquely involved in pathogenesis. A secreted cutinase [46, 47] (one of seven identified by Interpro in *F. graminearum*) and one other lipase critical for entry into plant cell walls was in this region, as well as a possible cerato-platanin that is implicated in plant cell death [48]. Another unexpected protein type was encoded by three secreted ribonuclease genes thought to be involved in nutrient scavenging after plant death [49]. Interestingly, as with all clustered regions shown in Supplementary Table 1, there were 72 proteins whose function is not known. While this may reflect a shortcoming of domains annotated in Interpro, it may be exploited in a search for the function of unknown and orphan genes using a "guilt by association" type of approach. While a region spanning roughly 10% of the genome and apparently covering most of an entire chromosome arm

might not seem likely to be cotranscribed all at the same time, it is possible that subregions are active. This organization would make it easier for transcription factors to encounter other genes whose expression is required during host colonization. Perhaps it is relevant that this region of the genome encodes several groups of transcription factors containing nuclear localization signals. It is, however, difficult to connect individual transcription factors to a particular process.

Another intriguing region in the *F. graminearum* genome was quite large, although much smaller than the one described above. This clustered area on supercontig 3.4 contained 345 total genes and 34 extracellular targeted genes (p-value 7.4e-3), as shown in Figure 8. Similar to the region discussed above, it included many pectin lyases, glycoside hydrolases, acetylases, oxidases and peptidases, essential components of the biomass degrading machinery. Another interesting feature of this region was the lipase gene, which might be key to piercing the waxy cuticle of plants. In yet another region of the *F. graminearum* genome, we found another putative cutinase gene clustered very closely with several peptidases, suggesting a role in plant pathogenesis. It is also important to note that all of these regions appeared to be somewhat conserved in the other two *Fusarium* genomes in our study, *Fusarium verticillioides* (three regions of 36, 33 and 2 target genes) and *Fusarium oxysporum* (three regions of 68, 80 and 2 target genes). The gene content, conservation and possible role in infection of these regions suggest further investigation in functional studies.

In another well studied genome, that of *Trichoderma atroviride*, there were several large clusters containing proteins similar to those found in *Fusarium graminearum*. One region in particular (shown in Figure 9) spanned only 24 target (of 136 genes total in the cluster)

genes and contained genes encoding proteases and glycoside hydrolases that are typically involved in biomass degradation (hypergeometric p-value 1.6e-5). Two differences, however, highlight this region as potentially important to *T. atroviride*'s role as a well known mycoparasite and biocontrol agent [50, 51]. Notable among the glycoside hydrolases (GH) in this cluster is a protein putatively annotated as belonging to family GH18 [45], a GH family well known for chitinase activity. Also, the cluster contained a gene encoding a protein with the domain "IPR015131:Killer toxin, Kp4", which in some fungi has been shown to be involved in killing other fungi [52].

We were interested in the conservation of clusters of extracellularly targeted genes. By and large the clusters appeared to contain genes producing biomass degrading proteins. In order to determine if the clusters contained the same genes, we collapsed the entire list of 3,291 clusters into a unique set by tabulating where the same Interpro annotations appeared in different clusters. We found that there were 1,451 unique gene clusters, containing different types of peptidases, representatives from different glycoside hydrolase families, different oxidases and other protein types, as would be expected by the wide variety of gene content in fungal genomes [15, 25, 29, 53]. We also excluded the effect of organization (order in which the genes appear) and only considered the Interpro annotations of the genes in the cluster. The regions that were identical across genomes were due to the number of unknown target proteins they encoded. A common cluster motif (see Supplementary Table 1) that was found in multiple genomes was one identifiable gene (peptidase or glycoside hydrolase) followed by a variable number of unknowns; this could be responsible for the low diversity seen in the heatmap (Figure 7a) and PCA analysis (above).

In contrast to the secretion clusters, clusters of genes that were nuclear targeted appeared to be more diverse in their function. For example, translation initiation factors appeared in clusters in 55 genomes, while glycoside hydrolases appeared in extracellular targeting clusters in 71 genomes (exact binomial test p-value 1.7e-8). Only the greatly reduced genome of *Encephalitozoon cuniculi* [54] did not appear to contain this type of feature. Additionally, 59 genomes contained some sort of kinase in clusters, while zinc fingers of various types were found in clusters in 66 genomes. This apparent diversity in types of genes clustered may be at least partially responsible for the high variation that was seen in the PCA analysis.

Clustering of Genes Involved in Biochemical Pathways

To determine how biochemical pathway clusters were distributed across the genomes, we proceeded with PCA analysis as described above. The genes involved in biochemical pathways appeared to have less variation than the previously discussed types of clusters. The first component contained 23% of the variation, but PC2 contained only 6%, and the first 10 principle components were necessary to span the first 50% of the variation in the data. This indicated that the distribution of clustering across the pathways was more diverse than in the previous analyses.

Investigation of the loadings in PC1 revealed that the biochemical pathways contributing to the highest diversity (top loadings in PC1) were largely due to particular pathways being clustered in only a few fungi. For example, from Table 4a, the pathway that ranked fourth in diversity was that of "PWY-6124 inosine 5 phosphate biosynthesis II", which is involved in the synthesis of purines. This pathway was clustered in the yeasts *Candida albicans* (both strains in our study, sc5314 and wo1) and *Debaryomyces hansenii*. There were several

126

categories in Table 4a that had a wider distribution across the fungi studied. On the top of the list, the "P221-PWY octane oxidation" pathway is involved in the breakdown of hydrocarbons, which can feed products into the beta oxidation pathway; portions of this pathway were clustered across 50 genomes. The highest percentage was found in the basidiomycete C*oprinus cinereus*, whose relatives are well known for degrading hydrocarbons [55]. However, one of the model basidiomycetes for bioremediation studies [55], *Phanerochaete chrysosporium*, had one of the lowest percentages of genes in this pathway clustered. Whether this indicates that the *C. cinereus* pathway is not as active as the *P. chrysosporium* pathway is not known; this would be a good subject for future studies.

We were interested in pathways that were clustered across all organisms, and analyzed this as described in the localization section above. We found that only one pathway, the "TRNA-CHARGING-PWY tRNA charging pathway", was clustered across all genomes. To identify other pathways that may be important in fungal biology, we plotted the distribution of occurrence of pathways across all genomes. This showed a group of eight pathways that had a significantly high occurrence (defined as 2 standard deviations above the mean of 20, data not shown). This list, shown in Table 4b, contained key pathways of cellular metabolism. Genes involved in fermentation, a type of anerobic respiration (identified by "ANARESP1-PWY respiration anaerobic"), are some of the genes involved in alcohol production. This highlights the central importance of such a process across fungi. Additionally, pathways involved in basic cellular energy metabolism appeared to be clustered across most genomes. These included gluconeogenesis and glycolysis (which are usually separated by two irreversible enzymes and likely share genes common to both pathways), as well as the tricarboxylic acid cycle (TCA). Interestingly "GLYOXYLATE-BYPASS glyoxylate cycle" is a pathway that allows organisms to use alternative carbon

sources, such as fatty acids and alcohols, and may be subject to repression until needed. While several of these pathways might be considered stress responses, the TCA cycle and glycolysis are likely to be active throughout cellular life. Still, it is possible that having some of the enzymes accessible at the same time might be advantageous to nuclear transcription machinery.

Paralogous Clusters of Genes in Fungal Genomes

A frequent question in genome evolution concerns the fate of duplicated genes [56]. In the context of genome organization, that question alters to "do duplicated genes in fungal genomes move to new locations or do they remain near their original copy?" Previous genome analysis has shown that fungal genomes can be very low in duplicated genes due to special mechanisms [57], or rich in large gene families [15, 58, 53]. However, outside of a few well studied gene families, not much is known about the organization of paralogous genes [11, 24, 25]. By using data from the Arvas et al. 2007 [32] study, we can gain insight into the fate of multiple gene families and their movements across the genome.

Arvas and colleagues ref [32] analyzed clusters of genes from fungal genomes that were grouped by protein sequence similarity with MCL [31]. Each group contained a unique ID that we were able to map onto the fungal genomes via the gene identifier. Our clustering algorithm was then used to find clusters of genes in fungal genomes that came from multiple gene families, which at times spanned multiple genomes. The Arvas et al. 2007 study produced 3,188 unique groups across 44 genomes, which due to time constraints could not be expanded to the full 72 genomes in this study.

General trends were presented above in the Results section. Our first question concerned what proportion of genes from multiple gene families remained clustered in the fungal genomes. From the data in Figure 4, the mean proportion of genes from multiple genomes was .36. Of these, the mean proportion of the genes from multiple gene families was .18, indicating that only a small minority of genes that were duplicated have drifted to new locations in the genome, far from the original copy (p-value 3.7 e-14). While the ages of the duplicated genes were not known, and may greatly affect this measure (older duplicates have a greater chance to move), the parameters of the MCL were created to make the groupings favor "recent" duplicates Arvas 2007 [32]. However, the data do not let us rule out the possibility that the duplicated genes in fungal genomes arose a long time ago, and subsequently moved.

The above result may lead one to conclude that the contribution of multiple gene families to genome structural evolution might be negligible. There were, however, many clusters of genes from multi-gene families that we discovered in the set of 44 fungal genomes. With the annotation from the MCL groups, we can investigate how the gene families were structured across the genomes. By using the PCA technique, as in the above analyses, we first found gene family clusters that were the most diverse in the genomes. We again chose to investigate the top 25 loadings in PC1, shown in Table 5. To gain insight into the functions of the genes, we mapped the Interpro annotations onto the MCL groups as they appeared in each genome, similar to the WoLF PSORT analysis above.

Some unusual trends emerged from a detailed analysis of the clusters in Table 5. The most diverse family in PC1 was a group of protein kinases found only in the biomass degrading basidiomycete fungi in the study, *Laccaria bicolor, Phanerochaete chrysosporium, Coprinus*

129

cinereus and *Postia placenta*. The span across basidiomycetes excluded the two basidiomycete plant pathogens, *Ustilago maydis* and *Puccinia graminis*. While *U. maydis* has the smallest basidiomycete genome (6,513 protein coding genes), and had one of the smallest proportions of genes in families (.24, ranking sixth from the bottom), which might cause a lack of clusterable duplicates, *P. graminis* has the second largest genome. In fact, of the MCL families listed in Table 5, there was no representation of gene clusters from either of the basidiomycete plant pathogens; instead they were found almost exclusively in the biomass degrading basidiomycetes, except for TribeFamily_85_t07_2.3, which identifies a key gene involved in secondary metabolite production, IPR006094:FAD linked oxidase containing genes [59]. Additionally, we note that in Table 5 many of the clusters of unknown genes were entirely in *L. bicolor*, the largest genome in the study, which likely accounts for its position in the hierarchical tree in Figure 11a and 11b.

We were also interested in identifying any gene families that were clustered across all genomes. Oddly, only one MCL group of protein kinases was clustered in all genomes. This group of kinases was apparently different enough from the ones above to form a separate MCL group. To find other genes that were clustered in many genomes, we sorted a list of occurrence of MCL groups across the genomes and chose an arbitrary cutoff of occurring in 30 genomes; this resulted in the 22 MCL groups shown in Table 5. It is interesting to note that mean number of MCL families clustered in the genomes was 2.27, which may be due to the number of multiple species from the same genus in the data set.

We mapped the functions as above to the gene families. Surprisingly, the top two clusters were protein kinases, which showed the dynamic nature of the gene type. Also of note was the lack of biomass degrading genes in the list. Only one, a peptidase family, was clustered

in 44 genomes. This likely underscores the specific nature of this type of gene and supports previous findings that biomass degrading glycoside hydrolases are not from recent duplications in at least two genomes, those of *Trichoderma reesei* [25] and *Aspergillus oryzae* [11]. Surprisingly, what we did find in this list was a large number of clustered transporters. Identified by IPR005828:General substrate transporter, IPR016196:Major facilitator superfamily, IPR005829:Sugar transporter, IPR001140:ABC transporter, transmembrane region and IPR002293:Amino acid/polyamine transporter I, transporters appeared in 9 of these top 22 clusters. Of these, it appears that most were sugar transporters. Why arrays of sugar transporters were clustered in the genome is unclear. However, sugar transport can be fairly specific [59], and it is plausible that several types may need to be deployed depending on the carbon source encountered in the environment.

CONCLUSIONS

We have successfully created a flexible algorithm for the discovery of genomic clusters of genes. The intent was to allow many different aspects of gene/protein annotation to be used as input. The operator only needs three pieces of data: the gene to annotation file, the gene to genome location file and a p-value cutoff. This allowed us to rapidly discover many types of clusters once entire genomes were annotated. The Perl script is available for download at kiddomics.com/ and is distributed under the GNU Public License

(http://www.gnu.org/licenses/gpl.html).

Overall, the amount of clustering detected was rather low. Few genomes had gene clusters totaling more than 50% of the annotation type. Still, the level of clustering we did find was rather striking, in that all genomes contained some level of clustering, including that of *N. crassa.* which may be considered a poor substrate for clustering (since it contains few

multiple gene families). The observed level of clustering might highlight the importance of organization for those genes that were clustered. This potential importance was most noticeable in the clusters of secreted genes, which span many different protein families, such as clusters of proteases, glycoside hydrolases and secondary metabolism genes colocated in the genome. While many aspects of gene clustering remain unexplored, our work establishes the foundation for future studies.

It has been suggested that the purpose of clustering is coregulation or cotranscription [6]. However, it is not entirely clear that clustered genes do indeed cotranscribe. The primary literature contains examples that appear to support both sides of the argument [10, 25, 60]. This indicates that we probably have not identified all factors that tie organization to transcription. Factors that affect formation and maintenance of genome clusters might also be factors that affect transcription, and thus work in our laboratory is ongoing to provide insights into the history of the genome clusters and factors that maintain cluster cohesiveness. One relevant aspect is the effect of lateral gene transfer, which has been suggested as important for bacterial operons [61].

By demonstrating that clustering occurs in an entire kingdom, we have shown that the organization of genes across genomes is an important feature that should be considered in studies of genome evolution. Indeed, few such studies have been performed, with this one being the broadest. Additional work will be needed to incorporate the many genomes still to be released. At last count, there were approaching 400 completed fungal genomes (www.genomesonline.org/), and with next generation technologies now being used for *de novo* sequencing, there will be ample substrate for future hypothesis testing [62].

MATERIALS AND METHODS

False Discovery Rate Estimation

In large data sets with thousands of statistical tests, it is possible that a portion of the statistical tests will produce false positives. There are two ways of dealing with this problem. The first is to throw away potentially false positive results, while the second is to estimate the potentially false positive portion of the results, called false discovery rate (FDR) estimation. The latter approach has become more common in recent years, as it is seen as a trade off between being too prudent, as in the Benjamini-Hochberg method or the Bonferroni correction, and allowing too many errors in the data set [63].

Currently there are several off the shelf methods for estimating the FDR. However, due to the statistical dependencies and the unique concave distribution of our results (data not shown), none of these were applicable. Therefore, a null model was constructed that was specific to the analysis that we performed. We developed a bootstrapping method, which is usually considered a conservative estimate of error; related estimations have been used for similar models involving the hypergeometric distribution [64]. We therefore defined the FDR for this clustering analysis as the probability that a gene can fall into the same cluster as the real cluster when the genome is randomized. For example, if gene 1 from genome A was in a cluster of genes that all contained the domain IPR000917:Sulfatase (using the actual genome sequence), then 150 randomizations were performed to find if the gene reformed that particular cluster. All genes and all annotations were considered. However, due to time constraints, we chose a subset of data across both phylogenetic and genome size diversity to produce an estimate, shown in Table 7.

Due to time constraints we sampled the genomes in our study according to phylogenetic placement and genome size, settling on 18 genomes (¼) for FDR estimation. In Table 5 we see that the mean FDR per category ranged between .07 and .05. This is important, as this likely reflects the ceiling (due to the conservative nature of bootstrapping) for error, and shows that our method is sound. There were some genomes that had a rather high FDR, such as *Coprinus cinereus,* which had a maximum of .15 FDR. This appeared to be due to several large duplications in the genome, and may be seen as an outlier that is not common in the data set.

Interpro Annotation

All genomes were downloaded from the locations listed in Table 1. The proteins were run on a 100 node power-pc cluster at the Center for Advanced Research Computing (http://www.hpc.unm.edu/). We used the most recent runInterproScan version available at the time, version 4.4, with database version 14. Child-to-parent mappings were performed using the provided map file from ftp://ftp.ebi.ac.uk/pub/databases/interpro/ and a custom Perl script to parse and format the data.

Biochemical Pathways Annotation

We downloaded pre-annotated genomes from the KEGG [20] database. There were 20 genomes in the study available that contained enzyme commission (EC) code annotations. To transfer EC codes to the remaining genomes, we performed BLAST [65] alignments on the entire fungal proteome (747,893 proteins total) and then calculated mutual best hit (MBH) clusters. We then used a majority vote (45% minimum) strategy to transfer EC codes from proteins that were downloaded with the EC codes to the proteins previously lacking EC codes. By adjusting the e-value and manually verifying the accuracy of the transfer on a

random sample of MBH clusters, we used a BLAST e-value cut off of e-120 as maximum for a protein to be in a MBH cluster. The vote strategy maybe seen as low; however, this was due to numerous proteins with incomplete EC codes (1.1.1.-, for example) that were placed into the groupings.

We then used a custom Perl script to convert the EC code associated files to a rudimentary Genbank format that was used as input to Pathologic [28]. The Pathologic tool automatically associates proteins with EC codes to pathways. While Pathologic can fill in missing enzymes with a "hole filler" tool, we did not apply the hole filler tool to the data set, as it would be outside our rather stringent quality control system.

Protein Targeting Annotation

Several software tools that identify protein localization signals are available. However, WoLF-PSORT   [26] provides the most complete annotation, providing nuclear, transmembrane, extracellular, mitochondrial, Golgi, vacuolar, peroxisomal, lysosomal, endoplasmic reticulum, cytoskeleton and cytosol targeting predictions. In brief, WoLF PSORT uses a K-nearest neighbor method of annotation in combination with sequence prediction methods. Manual inspection of the profiles of nearest neighbors revealed that there was considerable disagreement in the targeting of the nearest neighbors. To reduce the false assignment rate, results of Phobius [27] were compared to those of WoLF PSORT in the categories that Phobius provides, transmembrane and extracellular. It was evident that for both categories, Phobius and WoLF PSORT agreed on the location when the WoLF PSORT nearest neighbors were 65% or greater of the total neighbors given. This was used as the cutoff for all categories. We would like to point out that while WoLF PSORT provides

many locations for protein trafficking, only the six locations shown in **Figure 7a** were returned by the software.

TrkNClusterViz, the Genome Cluster Browser

In order to visualize the data in our study, we constructed a custom viewer using the Java language. In addition, the graphical interface was built using BioJava version 1.7.1 [66]. The intent of the viewer was to provide a graphical interface in which the user could view the genes in the genome, gene annotations for the genes as well as the clusters and annotations for the clusters. The required input files for viewing are a map format file for the genome, an annotation file (containing some sort of gene annotation), and as many as 4 cluster files. The map files, annotation files and viewer code used in our study will be available at kiddomics.com upon publication. All cluster files will be available as Supplementary file 1.

```
n"=# of annotated genes in scaffold

k'=0
k=0
FOREACH scaffold n" > 1
  clusterinfo = empty list;
  GET   n_scaff genes  sorted by increasing base position;
  For i=1:n_scaff do
 IF (i is annotated)
        k'++;
        IF (k' > 2)
             pvalue = hypergeometric(k',k,n,n');
             IF(pvalue < .05)
                  add i to clusterinfo
             ELSEIF(k' >= 3)
                  PRINT clusterinfo
                  k'=1
                  trim back the cluster to last good position;
               ELSE#not significant and cannot be trimmed;
                  k'=1;
                  k = 0;
   ELSIF(end of scaffold)
        PRINT cluster info
   ELSE
        k++;
```

Box 1. Pseudocode of the gene cluster search algorithm.
The variables n, n', k and k' are as described in the text.

Table 1. List of genomes used in the studies. Column 2 indicates the unique ID code used in Figures 1-5, Supplementary Table 2 and supplementary data files.

| Name	Code	Phylum	Source
Aspergillus niger	apsnig-na	Ascomycota	Joint Genome Institute, USA
Ashbya gossypii	ashgos-982	Ascomycota	Integr8, EBI
Aspergillus clavatus	aspcla-na	Ascomycota	Integr8, EBI
Aspergillus flavus	aspfla-na	Ascomycota	Broad Institute, USA
Aspergillus fumigatus	aspfum-na	Ascomycota	Integr8, EBI
Aspergillus nidulans	aspnid-na	Ascomycota	Broad Institute, USA
Aspergillus niger	aspnig-CBS51388	ascomycota	Integr8, EBI
Aspergillus oryzae	aspory-na	Ascomycota	Integr8, EBI
Aspergillus terreus	aspter-na	Ascomycota	Integr8, EBI
Batrachochytrium dendrobatidis	batden-na	Chytridiomycota	Broad Institute, USA
Botrytis cinerea	botcin-na	Ascomycota	Broad Institute, USA
candida_albicans_sc5314_assembly_21_1	canalb-sc5314	Ascomycota	Broad Institute, USA
candida_albicans_wo1_1	canalb-wol1	Ascomycota	Broad Institute, USA
Candida glabrata	cangla-na	Ascomycota	Broad Institute, USA
Candida guiliermondii	cangui-na	Ascomycota	Broad Institute, USA
Candida lusitaniae	canlus-na	Ascomycota	Broad Institute, USA
Candida parapsilosis	canpar-na	ascomycota	Broad Institute, USA
Candida tropicalis	cantro-na	Ascomycota	Broad Institute, USA
Chaetomium globosum	chaglo-na	Ascomycota	Integr8, EBI
Cochlibolus heterostrophus C5	cochet-C5	Ascomycota	Broad Institute, USA
Coccidioides immitis	cocimm-h53841	Ascomycota	Broad Institute, USA
Coccidioides immitis RMSCC 2394	cocimm-rmscc23941	Ascomycota	Broad Institute, USA
Coccodioides immitis RMSCC 3703	cocimm-rmscc37031	Ascomycota	Broad Institute, USA
Coccidioides immitis RS	cocimm-rs	Ascomycota	Broad Institute, USA
Coccidioides posadasii RMSCC 3488	cocpos-rmscc34881	Ascomycota	Broad Institute, USA
Coccidioides posadasii Silveira	cocpos-silveira	Ascomycota	Broad Institute, USA
Coprinus cinereus	copcin-okayama7#1302	Basidiomycota	Broad Institute, USA
Cryptococcus neoformans	cryneo-h99	Basidiomycota	Integr8, EBI
Cryphonectria parasitica	crypar-na	Ascomycota	Broad Institute, USA
Debaryomyces hansenii	debhan-na	Ascomycota	Integr8, EBI
Encephalitozoon cuniculi	enccun-na	Microsporidia	Integr8, EBI
Fusarium graminearum	fusgra-na	Ascomycota	Broad Institute, USA
Fusarium oxysporum	fusoxy-lycopersici	Ascomycota	Broad Institute, USA
Fusarium verticillioides	fusver-na	Ascomycota	Broad Institute, USA
Histoplasma capsulatum	hiscap-nam1	Ascomycota	Broad Institute, USA
Kluyveromyces lactis	klulac-na	Ascomycota	Integr8, EBI
Laccaria bicolor	lacbic-na	Basidiomycota	Joint Genome Institute, USA
Lodderomyces elongisporus	lodelo-na	Ascomycota	Integr8, EBI
Magnaporthe grisea	maggri-7015	Ascomycota	Integr8, EBI
Microsporum gypseum cbs 1118893_1	micgyp-cbs1188931	Ascomycota	Broad Institute, USA
Mycosphaerella fijiensis	mycfij-na	Ascomycota	Joint Genome Institute, USA
Mycosphaerella graminicola	mycgra-na	Ascomycota	Joint Genome Institute, USA
Nectria haematococca	nechae-na	Ascomycota	Joint Genome Institute, USA
Neosartorya fischeri	neofis-na	Ascomycota	Integr8, EBI
Neurospora crassa	neucra-na	Ascomycota	Broad Institute, USA
Neurospora tetrasperma	neutet-na	Ascomycota	Broad Institute, USA
Paracoccidioides brasiliensis PB01	parbra-pb01	Ascomycota	Broad Institute, USA
Phanerochaete chrysosporium	phachr-na	Basidiomycota	Joint Genome Institute, USA
Phycomyces blakesleeanus	phybla-na	Zygomycota	Joint Genome Institute, USA
Pichia stipitis	picsti-na	Ascomycota	Integr8, EBI
Postia placenta	pospla-na	Basidiomycota	Joint Genome Institute, USA
Puccinia graminis	pucgra-tritici2	Basidiomycota	Broad Institute, USA
Pyrehophora tritici repentis 1	pyrtri-repentis1	Ascomycota	Broad Institute, USA
Rhizopus oryzae	rhiory-na	Zygomycota	Broad Institute, USA
Saccharomyces cerevisiae	saccer-ATCC_204508	Ascomycota	Integr8, EBI
Saccharomyces cerevisiae RM11-1a	saccer-rm111a1	Ascomycota	Broad Institute, USA
S. japonicus yFS275	schjap-yfs275	Ascomycota	Broad Institute, USA
Schizosaccharomyces pombe	schpom-na	Ascomycota	Integr8, EBI
Sclerotinia sclerotiorum	sclscl-na	Ascomycota	Broad Institute, USA
Sporobolomyces roseus	sporos-na	Basidiomycota	Joint Genome Institute, USA
Sporotrichum thermophile	spothe-na	Ascomycota	Joint Genome Institute, USA
Stagonospora nodorum 1	stanod-na	Ascomycota	Broad Institute, USA
Thielavia terrestris NRRL 8126	thiter-NRRL8126	Ascomycota	Joint Genome Institute, USA
Trichoerma atroviride	triatr-na	Ascomycota	Joint Genome Institute, USA
Trichoderma reesei	triree-na	Ascomycota	Joint Genome Institute, USA
Trichoderma virens	trivir-na	Ascomycota	Joint Genome Institute, USA
Uncinocarpus reesii	uncree-na	Ascomycota	Broad Institute, USA
Ustilago maydis	ustmay-na	Basidiomycota	Integr8, EBI
Verticillium albo-atrum VaMs.102	veralb-atrum	Ascomycota	Broad Institute, USA
Verticillium dahliae VdLs.17	verdah-na	Ascomycota	Broad Institute, USA
Yarrowia lipolytica	yarlip-na	Ascomycota	Integr8, EBI
schoct-yfs286	schoct-yfs286	Ascomycota	Broad Institute, USA
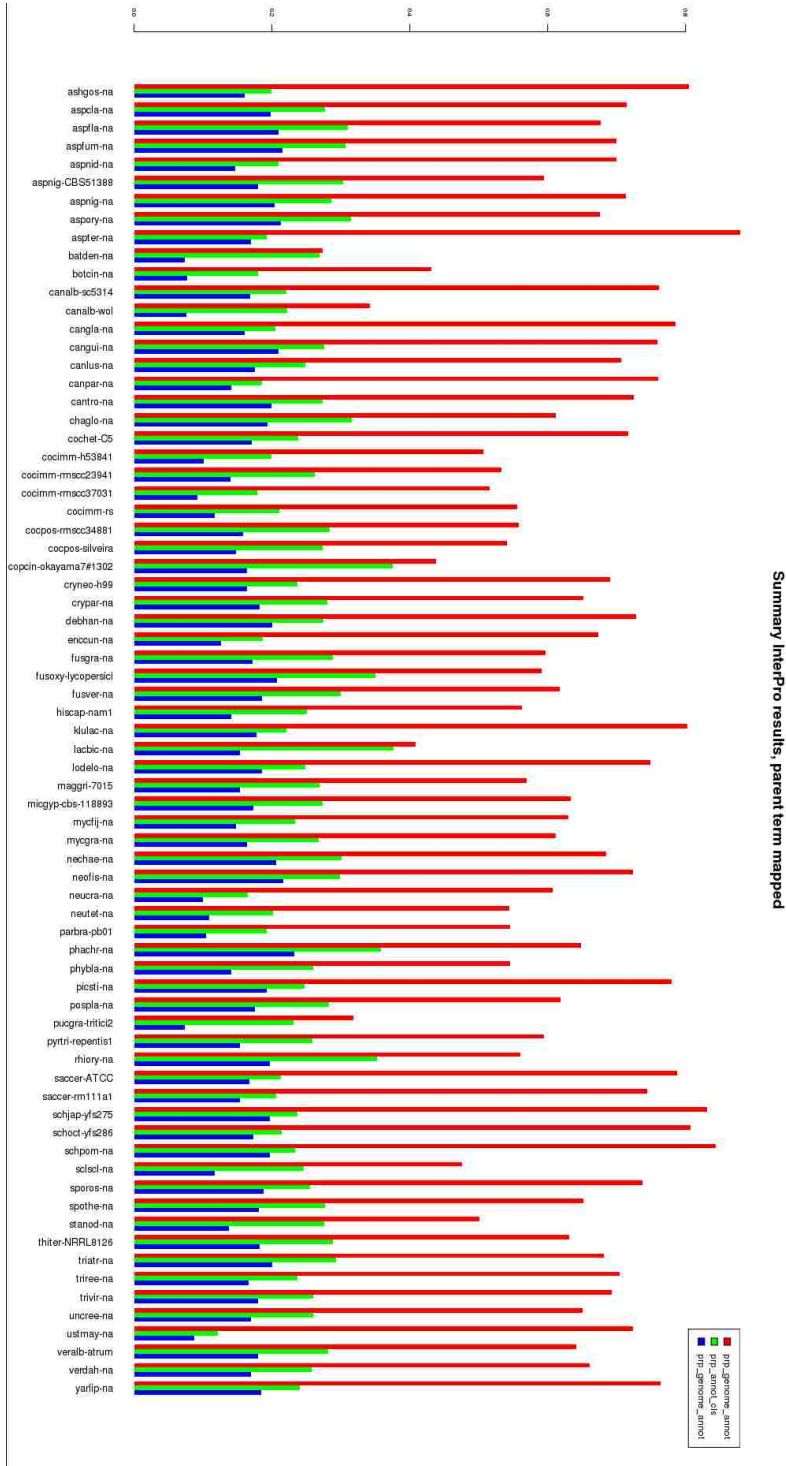
Figure 1. Summary of parent-term mapped Interpro [22] clusters. Red bars indicate the proportion of the genes in the genome that are annotated (contain at least one Interpro ID associated with the gene). Green bars indicate the proportion of annotated genes that are clustered, while blue bars indicate the proportion of all the genes in the genome that are clustered.
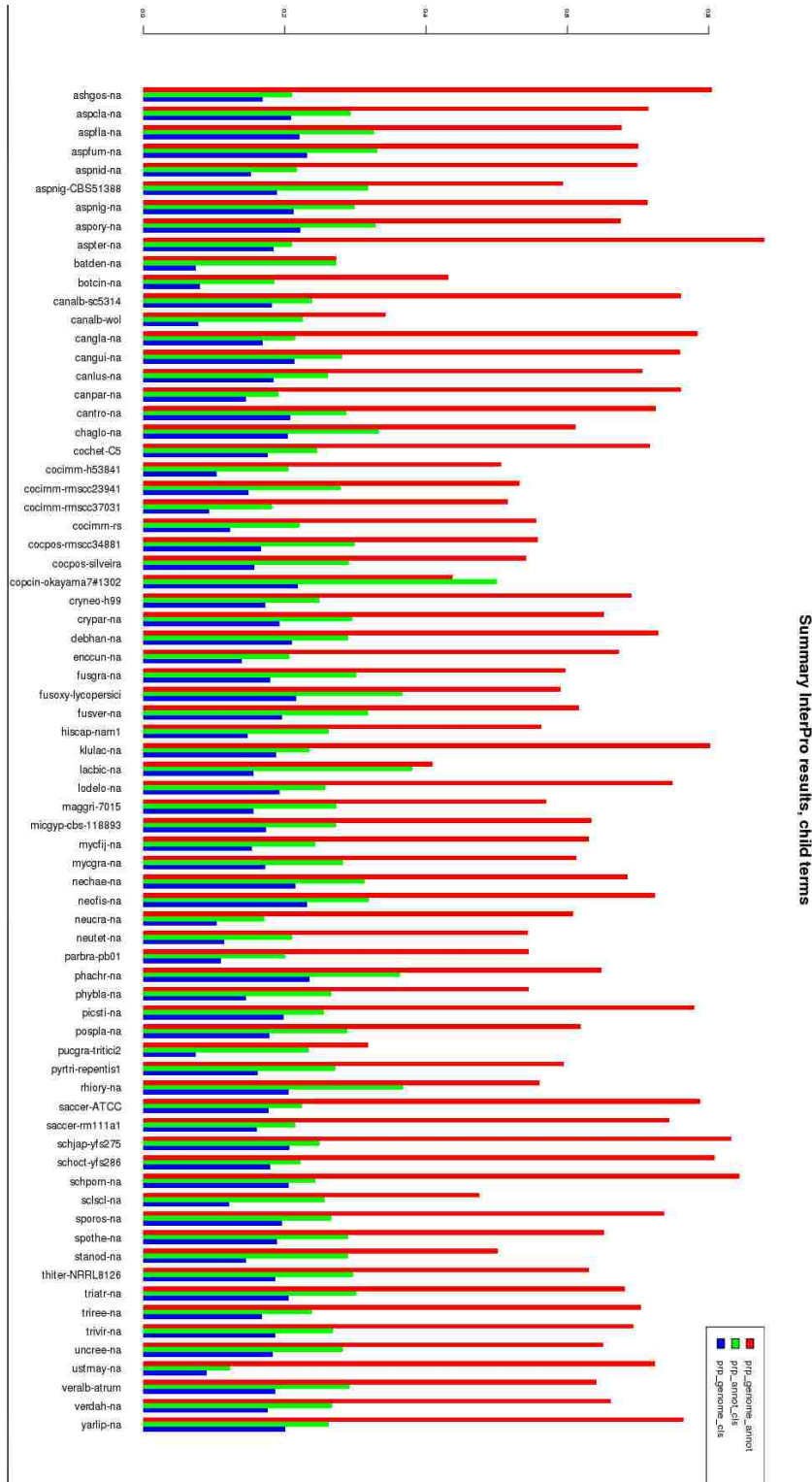
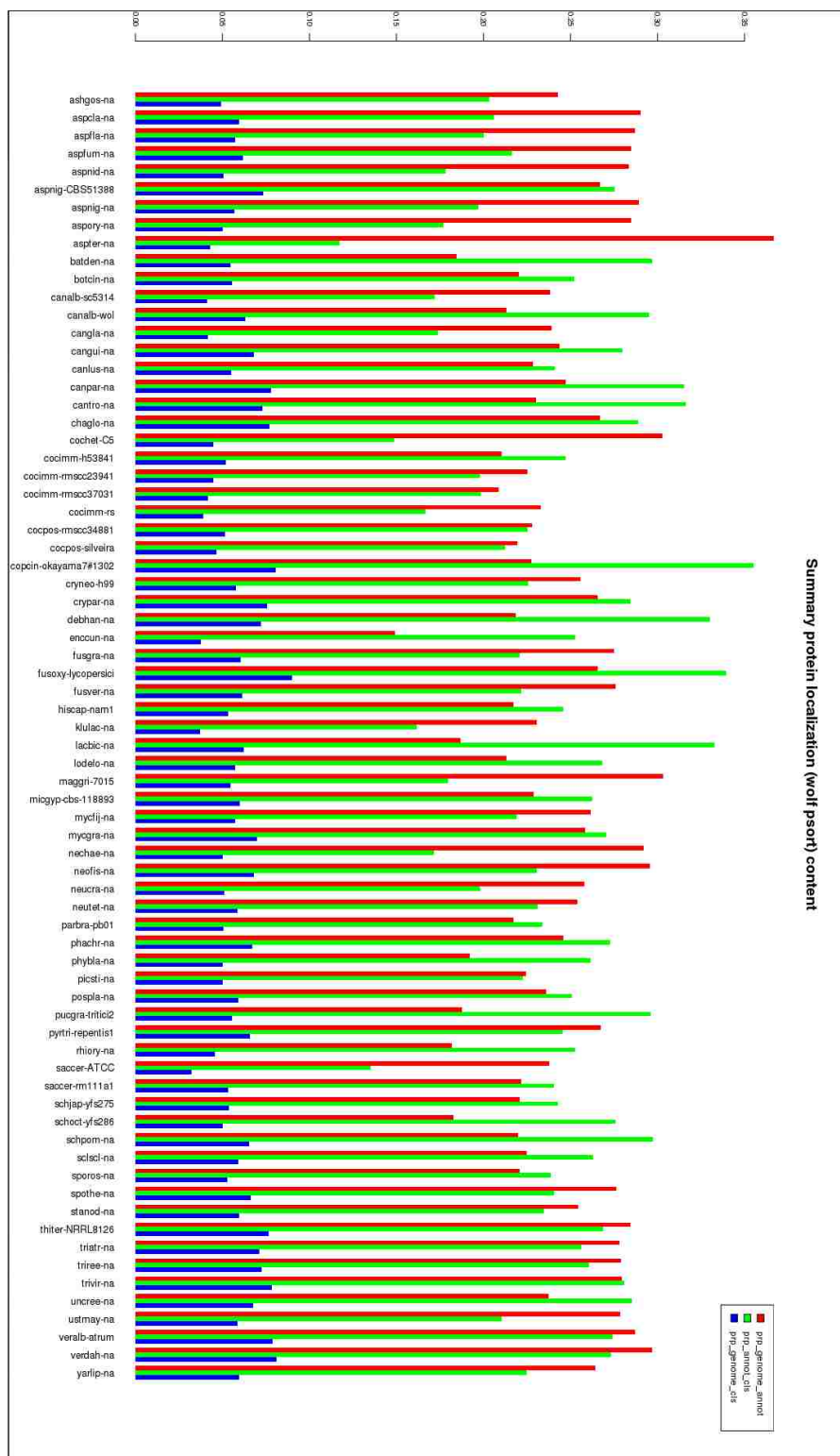**Figure 2.** Summary of child-term Interpro clusters. Colors are as in **Figure 1.**

Figure 3. Summary results of the clustering of genes annotated with WoLF-PSORT [26]. Colors are as in **Figure 1.**
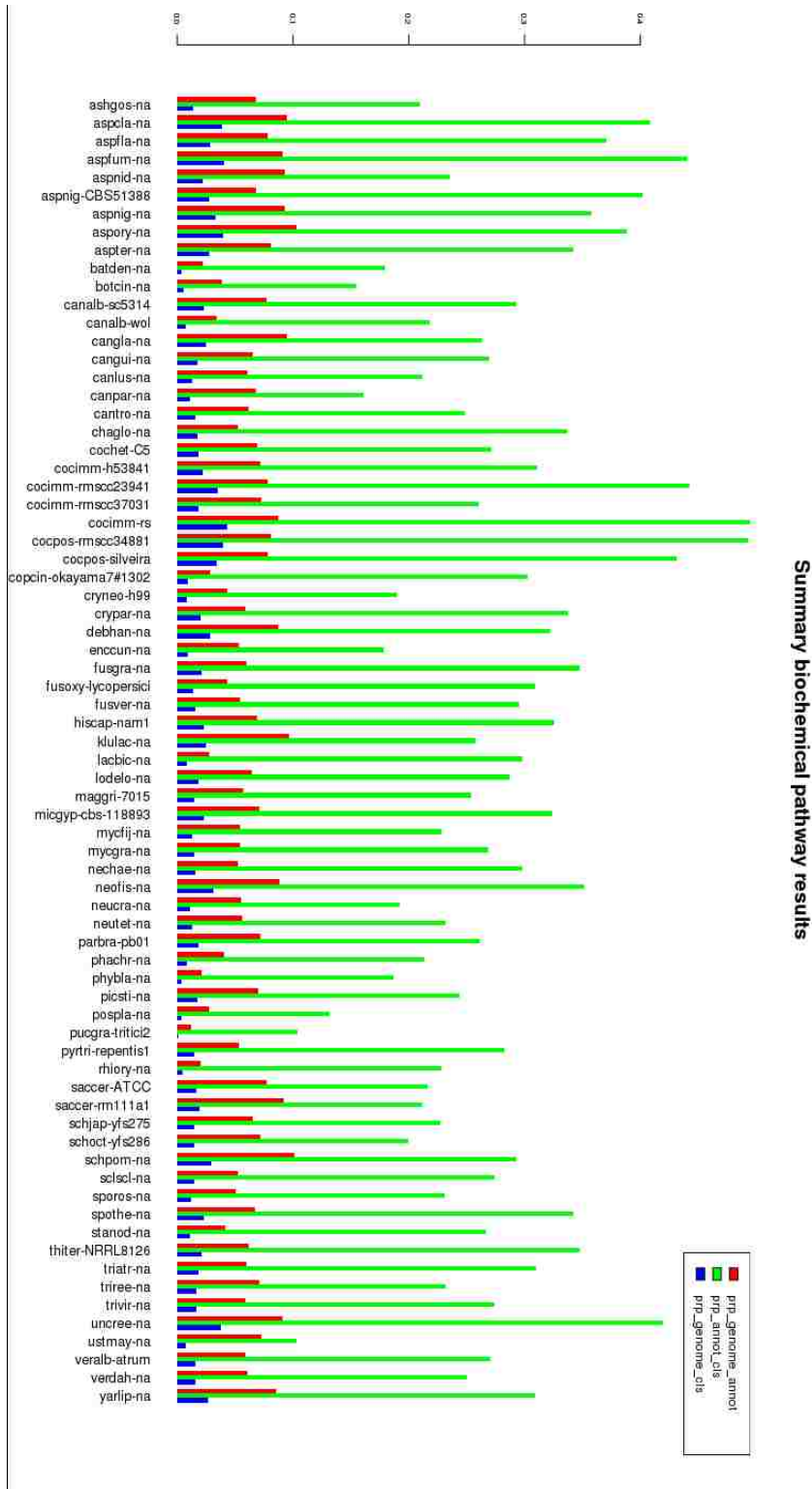
Figure 4a. Pathway dependent clusters of genes in genomes. Colors are as in **Figure 1.**
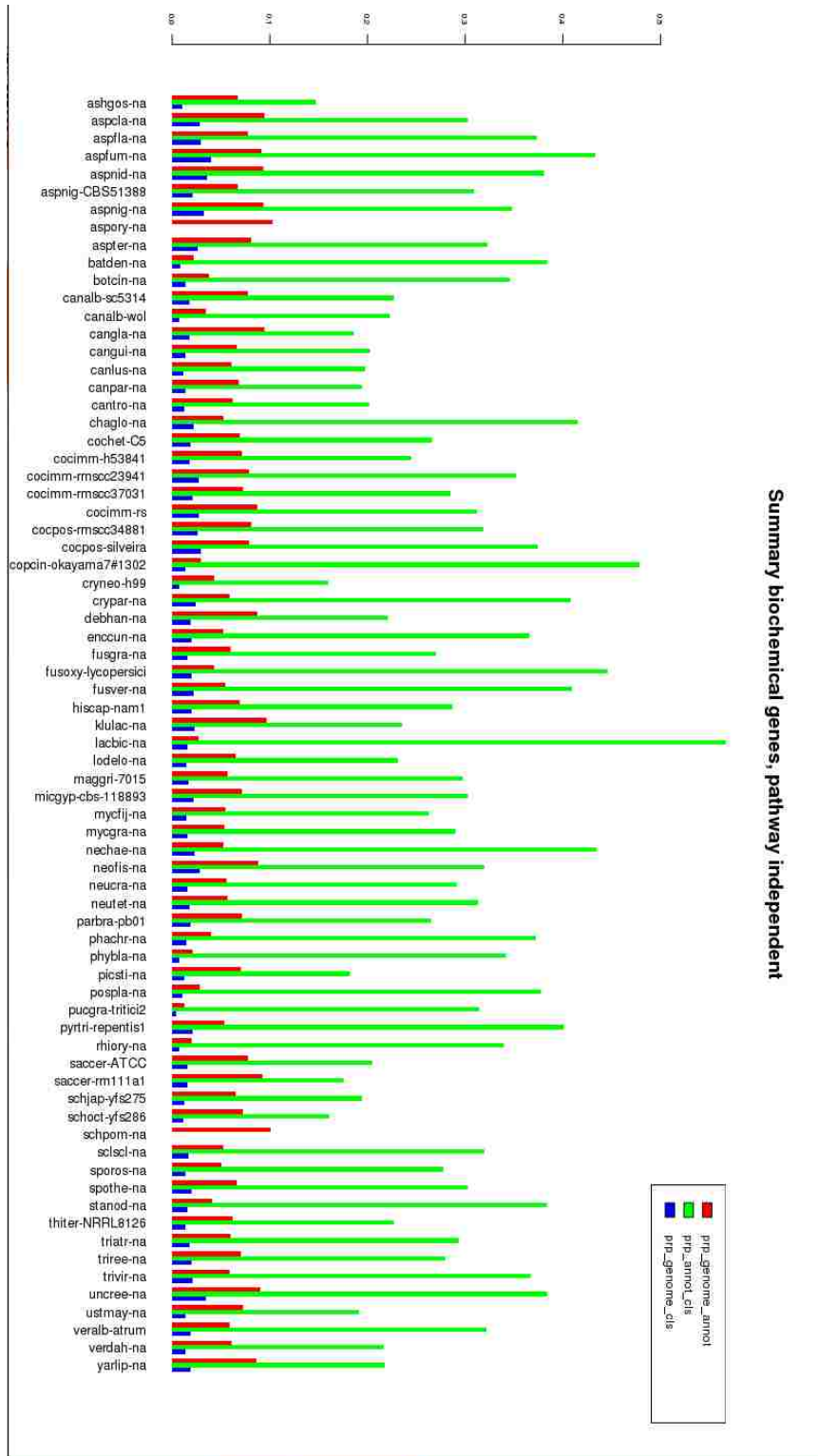
Figure 4b. Summary results of the pathway independent gene clusters in genomes. Colors are as in **Figure 1.**

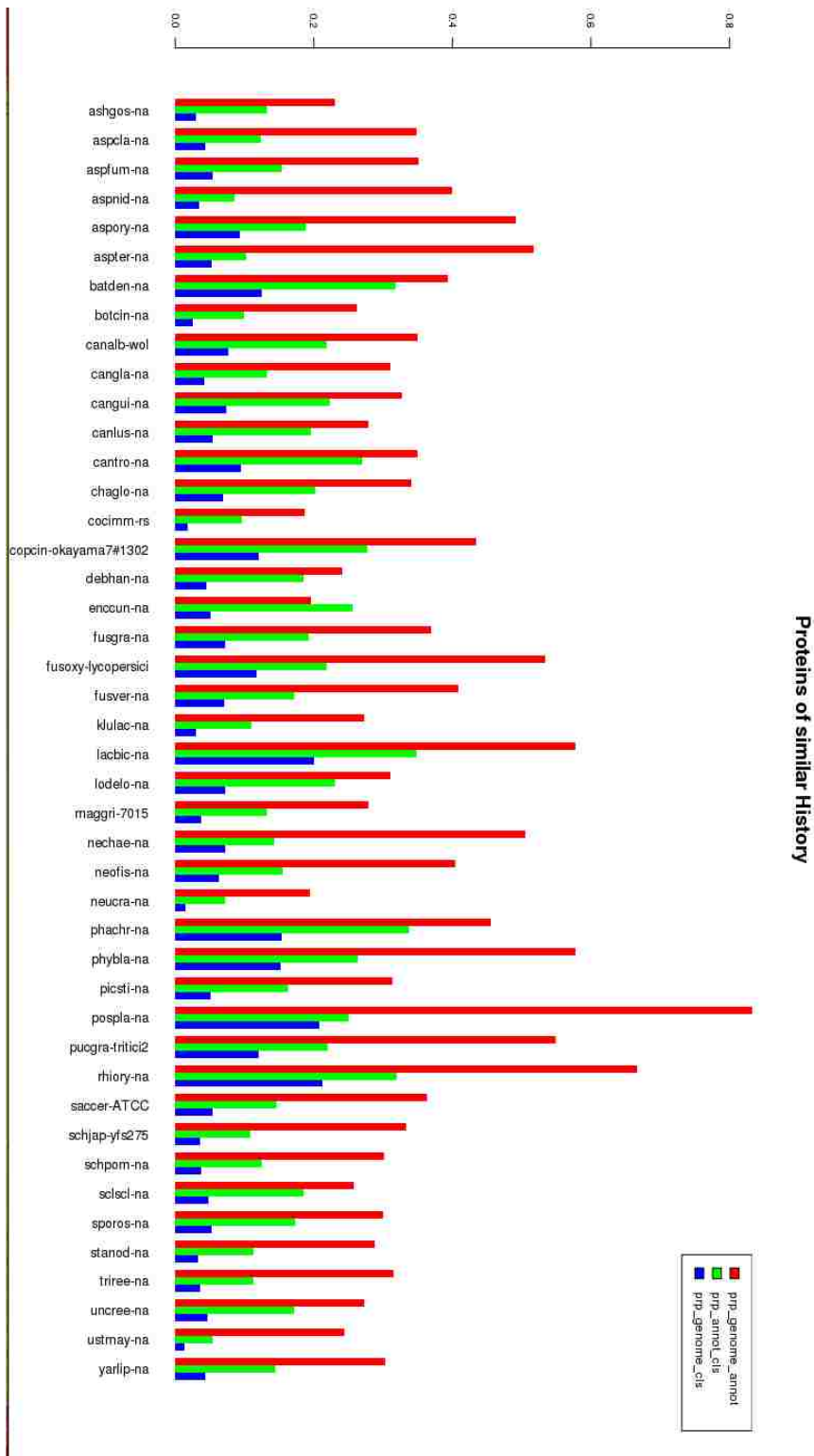**Figure 5.** Summary of the MCL [31] genome clusters. Colors are as in **Figure 1.**
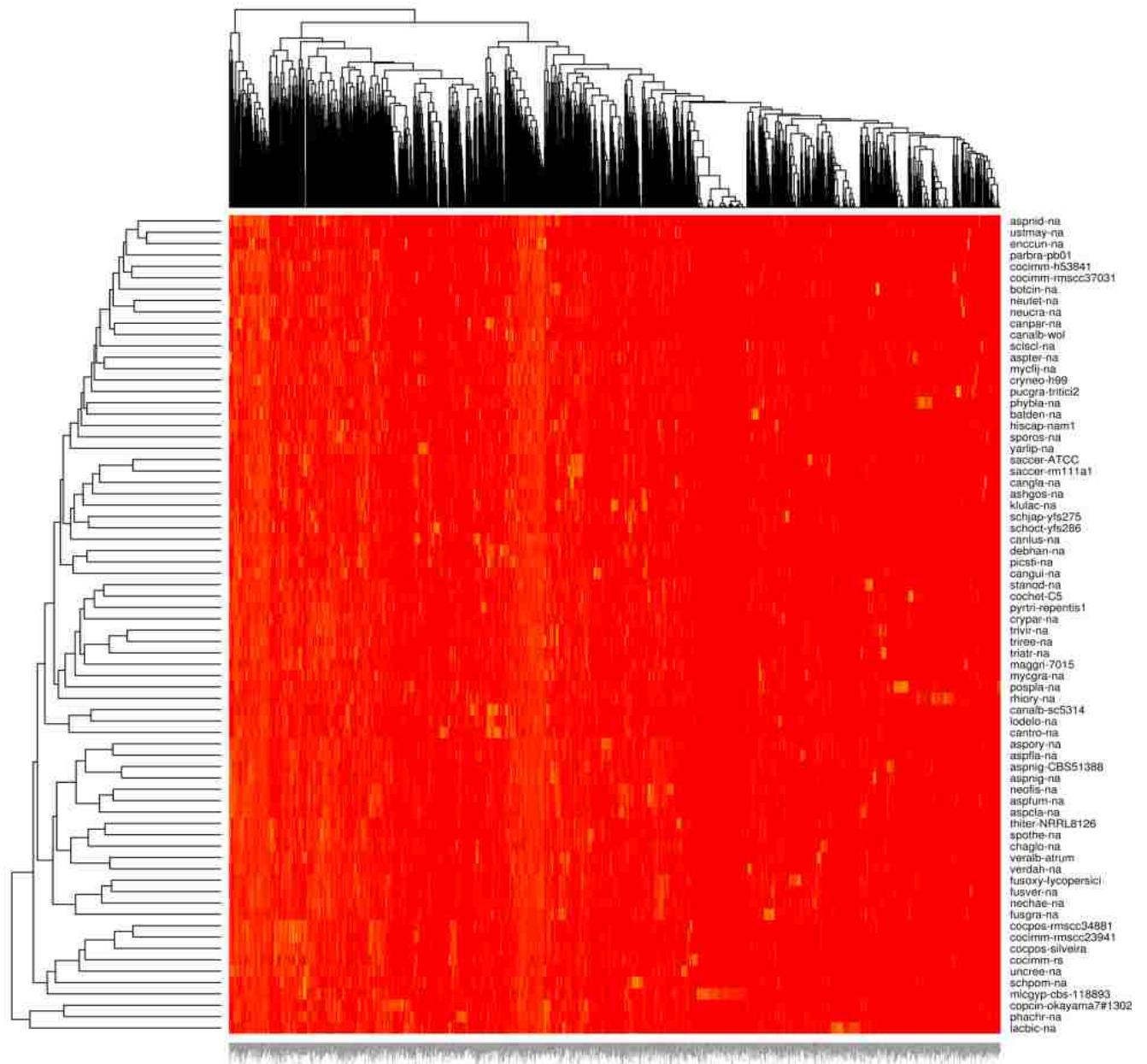
Figure 6a. Shows the heatmap results of the hierarchical clustering of the proportion of genes annotated to a particular category that were clustered in the genome. Red indicates a low proportion, and yellow indicates a high proportion. The x-axis shows the Interpro category, while the y-axis shows the genomes.

6b
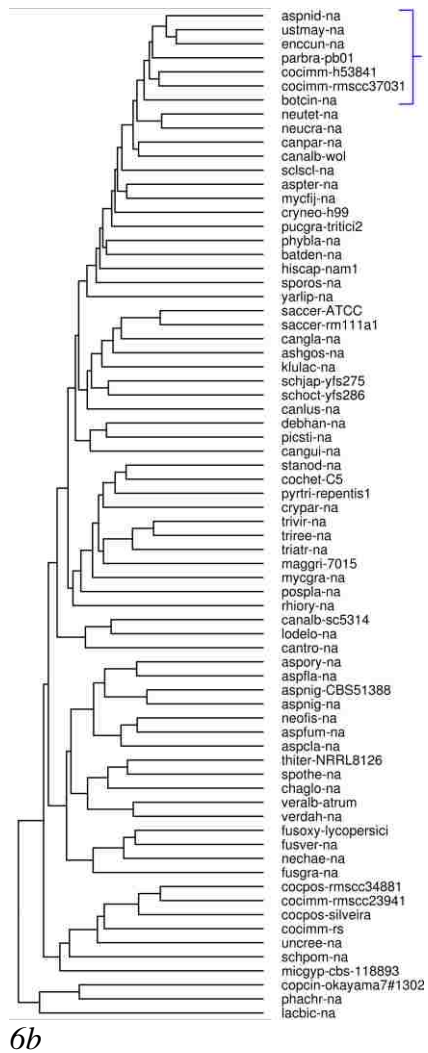
Figure 6b. The hierarchical tree from 6a showing the cluster profile relationships. The blue bracketed region contained several genomes that were displaced from their usual phylogenetic placement, consisting of plant and animal pathogens.

Table 2. Top loadings in principle component one of both the parent and child Interpro clusters. Column 2, labeled "role", shows the proposed general process category after manual review of the possible functions in which the family or domain is involved.

Interpro category	role
IPR013708.Shikimate.dehydrogenase.substrate.binding..N.terminal	primary metabolism
IPR016102.Succinyl.CoA.synthetase.like	primary metabolism
IPR004835.Fungal.chitin.synthase	growth and development
IPR011050.Pectin.lyase.fold.virulence.factor	biomass degradation
IPR009081.Acyl.carrier.protein.like	secondary metabolism
IPR008940.Protein.prenyltransferase	secondary metabolism
IPR002018.Carboxylesterase..type.B	biomass degradation
IPR000254.Cellulose.binding.region..fungal	biomass degradation
IPR009071.High.mobility.group..superfamily	mating type
IPR015815.3.hydroxyacid.dehydrogenase.reductase	biomass degradation
IPR000794.Beta.ketoacyl.synthase	secondary metabolism
IPR000172.Glucose.methanol.choline.oxidoreductase..N.terminal	biomass degradation
IPR008266.Tyrosine.protein.kinase..active.site	signaling
IPR012951.Berberine.berberine.like	secondary metabolism
IPR007867.Glucose.methanol.choline.oxidoreductase..C.terminal	biomass degradation
IPR008942.ENTH.VHS	protein interactions
IPR013830.Esterase..SGNH.hydrolase.type	biomass degradation
IPR001077.O.methyltransferase..family.2	regulation
IPR013094.Alpha.beta.hydrolase.fold.3	Interpro category since removed
IPR001242.Condensation.domain	secondary metabolism
IPR005024.Snf7	secretion
IPR016036.Malonyl.CoA.ACP.transacylase..ACP.binding	secondary metabolism
IPR013968.Polyketide.synthase..KR	secondary metabolism
IPR018392.Peptidoglycan.binding.lysin.domain	biomass degradation
IPR014031.Beta.ketoacyl.synthase..C.terminal	secondary metabolism
IPR006151.Quinate.shikimate.5.dehydrogenase.glutamyl.tRNA.reductase	primary metabolism
IPR006090.Acyl.CoA.oxidase.dehydrogenase..type.1	secondary metabolism
IPR012334.Pectin.lyase.fold	biomass degradation
IPR000910.High.mobility.group..HMG1.HMG2	mating type
IPR018205.VHS.subgroup	protein interactions
IPR006094.FAD.linked.oxidase..N.terminal	biomass degradation
IPR002014.VHS	protein interactions
IPR012132.Glucose.methanol.choline.oxidoreductase	biomass degradation
IPR002772.Glycoside.hydrolase..family.3..C.terminal	biomass degradation
IPR016166.FAD.binding..type.2	biomass degradation

Table 3. Interpro annotations that were clustered across all genomes. The mean proportion of genes from each particular category clustered in the fungal genomes is indicated.

| Interpro annotation	Mean proportion clustered
IPR017871.ABC.transporter..conserved.site	0.28
IPR011990.Tetratricopeptide.like.helical	0.28
IPR005225.Small.GTP.binding.protein	0.28
IPR011991.Winged.helix.repressor.DNA.binding	0.28
IPR001841.Zinc.finger..RING.type	0.30
IPR015880.Zinc.finger..C2H2.like	0.30
IPR012336.Thioredoxin.like.fold	0.30
IPR000504.RNA.recognition.motif..RNP.1	0.30
IPR014021.Helicase..superfamily.1.and.2..ATP.binding	0.30
IPR001650.DNA.RNA.helicase..C.terminal	0.31
IPR014729.Rossmann.like.alpha.beta.alpha.sandwich.fold	0.31
IPR012677.Nucleotide.binding..alpha.beta.plait	0.32
IPR013785.Aldolase.type.TIM.barrel	0.32
IPR015943.WD40.YVTN.repeat.like	0.32
IPR016024.Armadillo.type.fold	0.32
IPR001680.WD40.repeat	0.32
IPR003593.ATPase..AAA..type..core	0.33
IPR017441.Protein.kinase..ATP.binding.site	0.33
IPR008271.Serine.threonine.protein.kinase..active.site	0.34
IPR011009.Protein.kinase.like	0.34
IPR016135.Ubiquitin.conjugating.enzyme.RWD.like	0.36
IPR016027.Nucleic.acid.binding..OB.fold.like	0.36
IPR016196.Major.facilitator.superfamily..general.substrate.transporter	0.37
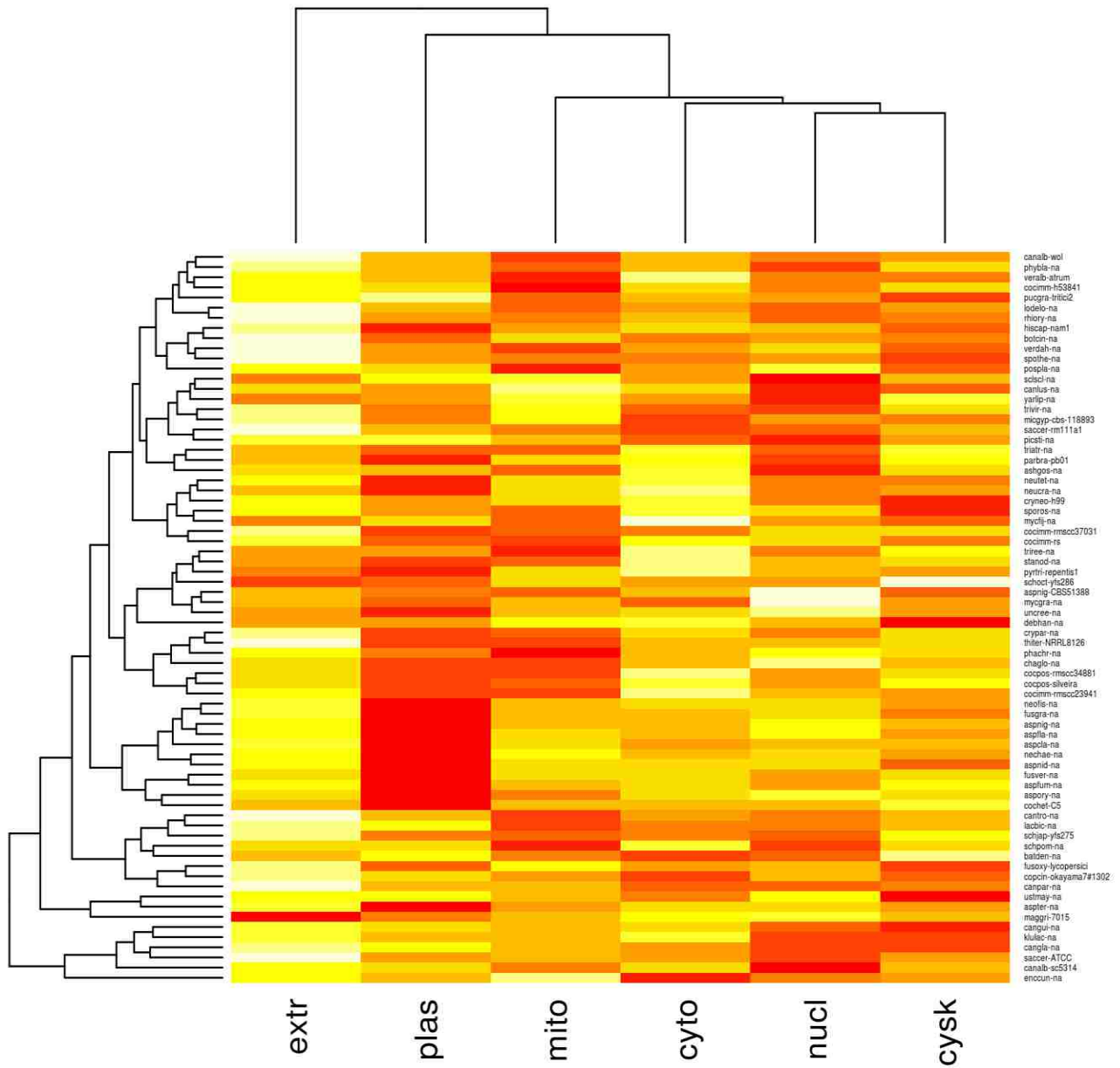
Figure 7a. The heatmap was constructed as in Figure 6a, except that the results of hierarchical clustering of genes annotated with WoLF PSORT are shown.
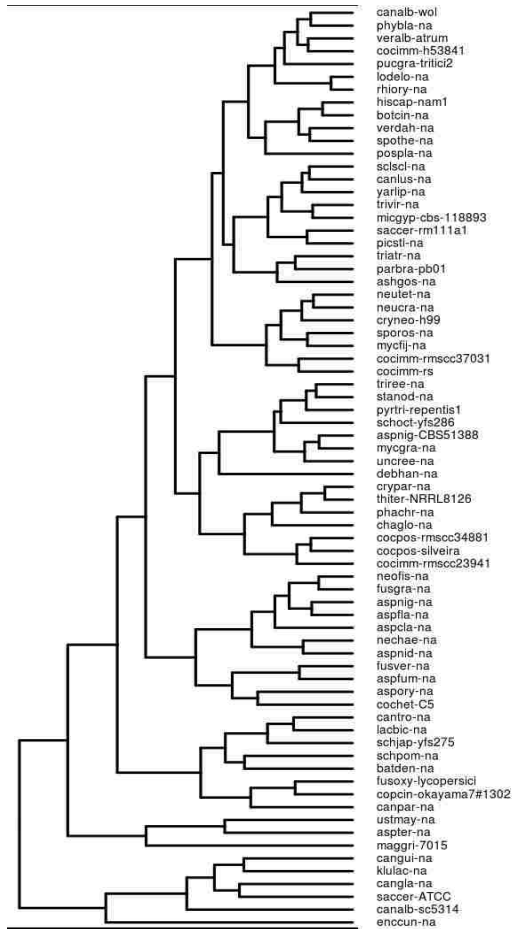
Figure 7b. Hierarchical clustering of genomes with similar proportions of genes targeted to a particular compartment that are clustered
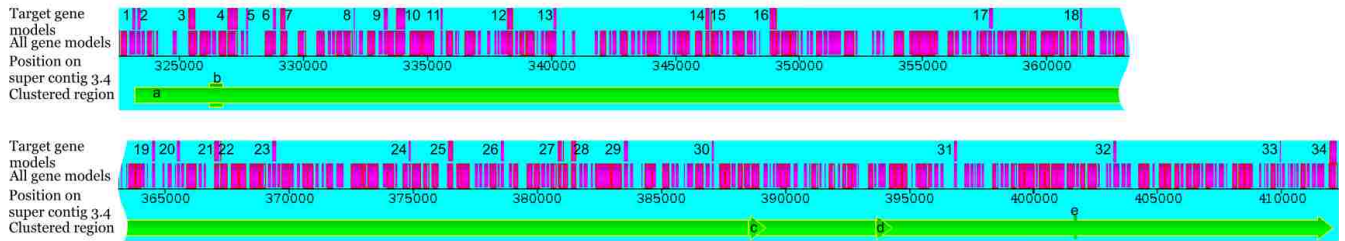
Figure 8. Region of the *Fusarium graminearum* genome containing many genes possibly involved in pathogenesis or biomass degradation as viewed in the TrkNClusterViz (see Materials and Methods).

The region contained a single cluster, but for clarity we have split the long region into two halves. The top row highlights the extracellular targeted genes, while the second row shows all genes. The green bars show the targeting cluster type. **a.** Extracellular targeted (all genes highlighted in the top row correspond to this cluster). **b.** Mitochondrial targeted. **c.** Nuclear targeted. **d.** Mitochondrial targeted. **e.** Nuclear targeted. Numbered genes were annotated as follows:**1.** fusgra-na_07350:IPR002018:Carboxylesterase, type B; **2.** fusgra-na_07351:IPR002016:Haem peroxidase, plant/fungal/bacterial; **3.** fusgra-na_07356:IPR005065:Platelet-activating factor acetylhydrolase, plasma/intracellular isoform II; **4.** fusgra-na_13137:ID_not_found; **5.** fusgra-na_07363:IPR002529:Fumarylacetoacetase, C-terminal-like; **6.** fusgra-na_07365:IPR001362:Glycoside hydrolase, family 32; **7.** fusgra-na_07366:IPR002509:Polysaccharide deacetylase; **8.** fusgra-na_13139:IPR002889:Carbohydrate-binding WSC; **9.** fusgra-na_07378:IPR010720:Alpha-L-arabinofuranosidase, C-terminal; **10.** fusgra-na_07380:IPR000073:Alpha/beta hydrolase fold-1; **11.** fusgra-na_07381:ID_not_found; **12.** fusgra-na_07384:ID_not_found; **13.** fusgra-na_07395:IPR002889:Carbohydrate-binding WSC; **14.** fusgra-na_07402:ID_not_found; **15.** fusgra-na_07421:ID_not_found; **16.** fusgra-na_13148:IPR000254:Cellulose-binding region, fungal; **17.** fusgra-na_07430:IPR004302:Chitin-binding, domain 3; **18.** fusgra-na_07462:IPR006076:FAD dependent oxidoreductase; **19.** fusgra-na_07475:IPR000726:Glycoside hydrolase, family 19, catalytic; **20.** fusgra-na_13156:ID_not_found; **21.** fusgra-na_07491:ID_not_found; **22.** fusgra-na_07495:IPR002889:Carbohydrate-binding WSC; **23.** fusgra-na_07502:IPR000209:Peptidase S8 and S53, subtilisin, kexin, sedolisin; **24.** fusgra-na_07523:IPR000782:FAS1 domain; **25.** fusgra-na_07528:ID_not_found; **26.** fusgra-na_07536:IPR001764:Glycoside hydrolase, family 3, N-terminal; **27.** fusgra-na_07543:IPR001952:Alkaline phosphatase; **28.** fusgra-na_07545:IPR006045:Cupin 1; **29.** fusgra-na_07550:IPR011050:Pectin lyase fold/virulence factor; **30.** fusgra-na_13164:IPR002198:Short-chain dehydrogenase/reductase SDR; **31.** fusgra-na_07601:ID_not_found; **32.** fusgra-na_13174:IPR001087:Lipase, GDSL; **33.** fusgra-na_13178:ID_not_found; **34.** fusgra-na_07664:ID_not_found.
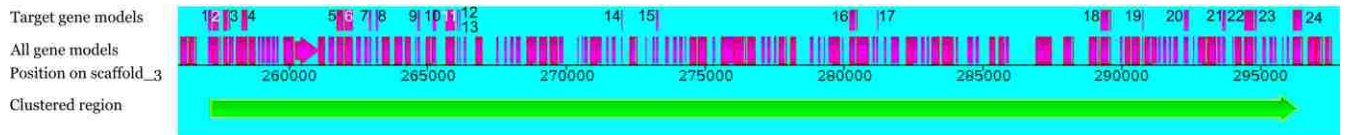
Figure 9. Region of the *Trichoderma atroviride* genome containing secreted genes that were statistically clustered.

The top row highlights the secreted genes. Gene annotations were:**1.** triatr-na_37404:IPR001910:Inosine/uridine-preferring nucleoside hydrolase; **2.** triatr-na_151059:IPR000560:Histidine acid phosphatase; **3.** triatr-na_36827:IPR006094:FAD linked oxidase, N-terminal; **4.** triatr-na_87477:IPR000408:Regulator of chromosome condensation, RCC1; **5.** triatr-na_36914:IPR000209:Peptidase S8 and S53, subtilisin, kexin, sedolisin; **6.** triatr-na_87489:ID_not_found:triatr-na_87489; **7.** triatr-na_138566:IPR011329:Killer toxin, Kp4/SMK-like, core; **8.** triatr-na_87493:ID_not_found:triatr-na_87493; **9.** triatr-na_127920:IPR008972:Cupredoxin; **10.** triatr-na_79492:IPR001223:Glycoside hydrolase, family 18, catalytic domain; **11.** triatr-na_35552:IPR001764:Glycoside hydrolase, family 3, N-terminal; **12.** triatr-na_36137:IPR008972:Cupredoxin; **13.** triatr-na_87502:ID_not_found:triatr-na_87502; **14.** triatr-na_19185:ID_not_found:triatr-na_19185; **15.** triatr-na_79507:IPR002053:Glycoside hydrolase, family 25; **16.** triatr-na_87546:IPR005151:Peptidase S41; **17.** triatr-na_87548:IPR011058:Cyanovirin-N; **18.** triatr-na_151132:IPR000322:Glycoside hydrolase, family 31; **19.** triatr-na_87574:ID_not_found:triatr-na_87574; **20.** triatr-na_87581:IPR000743:Glycoside hydrolase, family 28; **21.** triatr-na_36606:ID_not_found:triatr-na_36606; **22.** triatr-na_87590:IPR000254:Cellulose-binding region, fungal; **23.** triatr-na_36761:IPR000421:Coagulation factor 5/8 type, C-terminal; **24.** triatr-na_37346:IPR006102:Glycoside hydrolase family 2, immunoglobulin-like beta-sandwich.

Table 4a. Biochemical pathways that were the most diverse in clustering proportion across all genomes. The top 25 are shown for simplicity.

Interpro annotation	PC1 loadings
PWY.5084.2.ketoglutarate.dehydrogenase.complex	0.0257
P221.PWY.octane.oxidation	0.0201
COLANSYN.PWY.colanic.acid.building.blocks.biosynthesis	0.0149
PWY.6124.inosine.5..phosphate.biosynthesis.II	0.0133
PWY.3841.formylTHF.biosynthesis.II	0.0126
PWY0.162.pyrimidine.ribonucleotides..i.de.novo..i..biosynthesis	0.0118
PWY.5667.CDP.diacylglycerol.biosynthesis.I	0.0103
THIOREDOX.PWY.thioredoxin.pathway	0.0103
PWY.5123..i.trans.trans..i..farnesyl.diphosphate.biosynthesis	0.0100
PWY.5136.fatty.acid..beta..oxidation.II..core.pathway.	0.0099
ANAPHENOXI.PWY.phenylalanine.degradation.II..anaerobic.	0.0096
ALANINE.VALINESYN.PWY.alanine.biosynthesis.I	0.0070
PWY.66.GDP.L.fucose.biosynthesis.I..from.GDP.D.mannose.	0.0067
PWY.1881.formate.oxidation.to.CO.sub.2..sub.	0.0067
PWY0.1321.formate.to.nitrate.electron.transfer	0.0067
PWY0.1355.formate.to.trimethylamine.N.oxide.electron.transfer	0.0067
PWY0.1356.formate.to.dimethyl.sulfoxide.electron.transfer	0.0067
PWY4FS.8.phosphatidylglycerol.biosynthesis.II	0.0064
PWY3O.4106.NAD.salvage.pathway.III	0.0062
PWY.5194.siroheme.biosynthesis	0.0060
PWY.5695.urate.biosynthesis	0.0059
PWY.702.methionine.biosynthesis.II	0.0051
PWY.5941.glycogen.degradation.II	0.0050
PWY.4984.urea.cycle	0.0049
PWY0.1061.superpathway.of.alanine.biosynthesis	0.0046

Table 4b. Pathways that were found to be clustered across the highest number of genomes.

Pathway	No. of genomes containing pathway clustered
ANARESP1.PWY.respiration..anaerobic.	66
GLUCONEO.PWY.gluconeogenesis	66
GLYCOLYSIS.glycolysis.I	69
GLYOXYLATE.BYPASS.glyoxylate.cycle	66
P185.PWY.formaldehyde.assimilation.III..dihydroxyacetone.cycle.	67
PWY.561.superpathway.of.glyoxylate.cycle	69
PWY.5690.TCA.cycle.variation.III..eukaryotic.	68
TRNA.CHARGING.PWY.tRNA.charging.pathway	72

Table 5. False discovery rates for each genome and each annotation category. For genomes where zeros appear, no clusters were found in the randomized genomes. The abbreviated genome identifier for species name is shown in **Table 1**. The Mean indicated that there was a very low FDR in our study, given the conservative nature of bootstrapping tests.

Genome	parent mapped IPR	Child mapped IPR	biochemical pathways – dependent	biochemical pathways – independent	MCL
aspnid-na	0.05	0.05	0.04	0.08	0.02
aspory-na	0.09	0.09	0.09	0	0.05
botcin-na	0	0.04	0.02	0.08	0.02
canalb-wol	0.08	0.08	0.06	0.05	0.07
chaglo-na	0.1	0.1	0.09	0.09	0.07
cocimm-rmscc37031	0.04	0.04	0.04	0.05	0.04
copcin-okayama7#1302	0.11	0.15	0.08	0.11	0.09
fusgra-na	0.09	0.09	0.09	0.04	0.06
mycgra-na	0.08	0.08	0.06	0.07	0.08
neucra-na	0.04	0.04	0.03	0.07	0.02
phachr-na	0.09	0.09	0.05	0.09	0.08
phybla-na	0.06	0.06	0.03	0.08	0.07
rhiory-na	0.1	0.1	0.06	0.08	0.09
saccer-ATCC	0.07	5.36E-006	0.05	0.04	0.06
schpom-na	0	6.84E-006	0.08	0	0.05
triree-na	0.07	0.07	0.05	0.04	0.03
ustmay-na	0	0.02	0.01	0.03	0.01
veralb-atrum	0	0.08	0.06	0.08	NA
Mean	0.06	0.07	0.05	0.06	0.05

Figure 10a. The heatmap resulting from the hierarchical clustering of the MCL genome clusters.
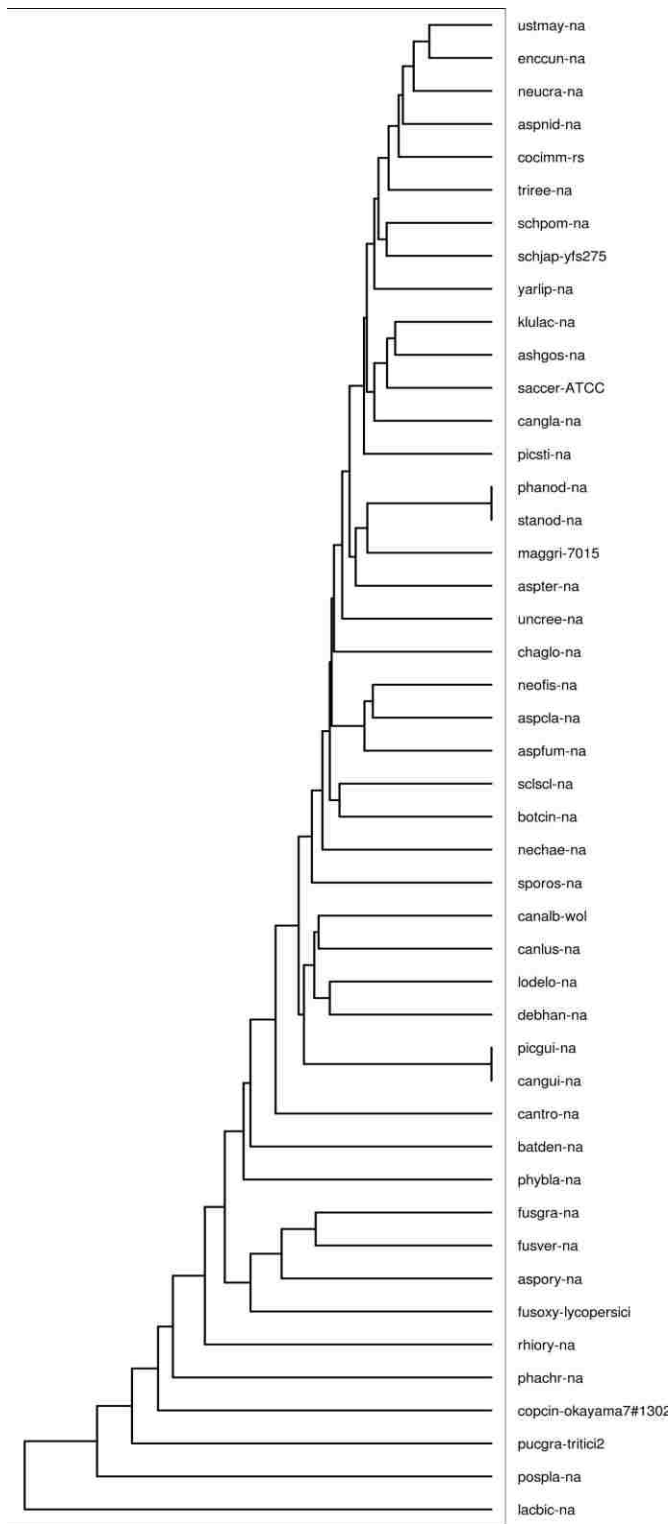
155

Figure 10b. Highlight of the resulting dendrogram of the hierarchical clustering in Figure 10a.

156

References

1.   Grigoriev IV, Martinez DA, Salamov AA: **Fungal Genome Annotation**. In *Bioinformatics*. 1st edition. edited by Arora DK, Berka R, Singh GB Amsterdam, the Netherlands: Elsevier Science; 2006, **6**:350.

2.   Francino MP: **An adaptive radiation model for the origin of new gene functions**. *Nat Genet* 2005, **37**:573-7.

3.   Kim J, Vanguri S, Boeke J, Gabriel A, Voytas D: **Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence**. *Genome Research* 1998, **8**:464-478.

4.   Daboussi M, Capy P: **Transposable elements in filamentous fungi**. *Annual Review of Microbiology* 2003, **57**:275-299.

5.   Samonte RV, Eichler EE: **Segmental duplications and the evolution of the primate genome**. *Nat Rev Genet* 2002, **3**:65-72.

6.   Hurst L, Pal C, Lercher M: **The evolutionary dynamics of eukaryotic gene order**. *Nat. Rev. Genet.* 2004, **5**:299-310.

7.   Wong S, Wolfe KH: **Birth of a metabolic gene cluster in yeast by adaptive gene relocation**. *Nat Genet* 2005, **37**:777-782.

8.   Lynch M: *The Origins of Genome Architecture*. 1st edition. Sinauer Associates Inc; 2007.

9.   Ohno S: *Evolution by gene duplication*. Springer-Verlag; 1970.

10.   Machida M, Asai K, Sano M, Tanaka T, Kumagai T, Terai G, Kusumoto K, Arima T, Akita O, Kashiwagi Y, Abe K, Gomi K, Horiuchi H, Kitamoto K, Kobayashi T, Takeuchi M, Denning D, Galagan J, Nierman W, Yu J, Archer D, Bennett J, Bhatnagar D, Cleveland T, Fedorova N, Gotoh O, Horikawa H, Hosoyama A, Ichinomiya M, Igarashi R, Iwashita K, Juvvadi P, Kato M, Kato Y, Kin T, Kokubun A, Maeda H, Maeyama N, Maruyama J, Nagasaki H, Nakajima T, Oda K, Okada K, Paulsen I, Sakamoto K, Sawano T, Takahashi M, Takase K, Terabayashi Y, Wortman J, Yamada O, Yamagata Y, Anazawa H, Hata Y, Koide Y, Komori T, Koyama Y, Minetoki T, Suharnan S, Tanaka A, Isono K, Kuhara S, Ogasawara N, Kikuchi H: **Genome sequencing and analysis of *Aspergillus oryzae***. *Nature* 2005, **438**:1157-1161.

11.   Khaldi N, Wolfe KH: **Elusive Origins of the Extra Genes in *Aspergillus oryzae***. *PLoS ONE* 2008, **3**:e3036.

12.   Doi R, Kosugi A: **Cellulosomes: Plant-cell-wall-degrading enzyme complexes**. *Nature Reviews Microbiology* 2004, **2**:541-551.

13. Fujino T, Beguin P, Aubert JP: **Organization of a *Clostridium thermocellum* gene cluster encoding the cellulosomal scaffolding protein CipA and a protein possibly involved in attachment of the cellulosome to the cell surface.** *J. Bacteriol.* 1993, **175**:1891-1899.

14. Moore RN, Bigam G, Chan JK, Hogg AM, Nakashima TT, Vederas JC: **Biosynthesis of the hypocholesterolemic agent mevinolin by *Aspergillus terreus*. Determination of the origin of carbon, hydrogen, and oxygen atoms by carbon-13 NMR and mass spectrometry**. *J. Am. Chem. Soc.* 1985, **107**:3694-3701.

15. Martinez D, Larrondo L, Putnam N, Gelpke M, Huang K, Chapman J, Helfenbein K, Ramaiya P, Detter J, Larimer F, Coutinho P, Henrissat B, Berka R, Cullen D, Rokhsar D: **Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78**. *Nature Biotechnology* 2004, **22**:695-700.

16. Cohen BA, Mitra RD, Hughes JD, Church GM: **A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression**. *Nat Genet* 2000, **26**:183-186.

17. Semon M, Duret L: **Evolutionary Origin and Maintenance of Coexpressed Gene Clusters in Mammals**. *Mol Biol Evol* 2006, **23**:1715-1723.

18. Lee JM, Sonnhammer EL: **Genomic Gene Clustering Analysis of Pathways in Eukaryotes**. *Genome Res.* 2003, **13**:875-882.

19. Yi G, Sze S, Thon MR: **Identifying clusters of functionally related genes in genomes**. *Bioinformatics* 2007, **23**:1053-1060.

20. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome**. *Nucleic Acids Research* 2004, **32**:D277-D280.

21. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis S, Matese J, Richardson J, Ringwald M, Rubin G, Sherlock G, Gene Ontology Consortium: **Gene Ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**:25-29.

22. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJA, Zdobnov E: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites**. *Nucleic Acids Research* 2001, **29**.

23. Dean R, Talbot N, Ebbole D, Farman M, Mitchell T, Orbach M, Thon M, Kulkarni R, Xu J, Pan H, Read N, Lee Y, Carbone I, Brown D, Oh Y, Donofrio N, Jeong J, Soanes D, Djonovic S, Kolomiets E, Rehmeyer C, Li W, Harding M, Kim S, Lebrun M, Bohnert H, Coughlan S, Butler J, Calvo S, Ma L, Nicol R, Purcell S, Nusbaum C,

Galagan J, Birren B: **The genome sequence of the rice blast fungus** *Magnaporthe grisea*. *Nature* 2005, **434**:980-986.

24. Fedorova N, Khaldi N, Joardar V, Maiti R, Amedeo P, Anderson M, Crabtree J, Silva J, Badger J, Albarraq A, Angiuoli S, Bussey H, Bowyer P, Cotty P, Dyer P, Egan A, Galens K, Fraser-Liggett C, Haas B, Inman J, Kent R, Lemieux S, Malavazi I, Orvis J, Roemer T, Ronning C, Sundaram J, Sutton G, Turner G, Venter J, White O, Whitty B, Youngman P, Wolfe K, Goldman G, Wortman J, Jiang B, Denning D, Nierman W: **Genomic Islands in the Pathogenic Filamentous Fungus** *Aspergillus fumigatus*. *PLoS Genet* 2008, **4**:e1000046.

25. Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, Chapman J, Chertkov O, Coutinho PM, Cullen D, Danchin EGJ, Grigoriev IV, Harris P, Jackson M, Kubicek CP, Han CS, Ho I, Larrondo LF, de Leon AL, Magnuson JK, Merino S, Misra M, Nelson B, Putnam N, Robbertse B, Salamov AA, Schmoll M, Terry A, Thayer N, Westerholm-Parvinen A, Schoch CL, Yao J, Barbote R, Nelson MA, Detter C, Bruce D, Kuske CR, Xie G, Richardson P, Rokhsar DS, Lucas SM, Rubin EM, Dunn-Coleman N, Ward M, Brettin TS: **Genome sequencing and analysis of the biomass-degrading fungus** *Trichoderma reesei* **(syn.** *Hypocrea jecorina***)**. *Nat Biotech* 2008, **26**:553-560.

26. Horton P, Park K, Obayashi T, Fujita N, Harada H, Adams-Collier C, Nakai K: **WoLF PSORT: protein localization predictor**. *Nucl. Acids Res.* 2007:gkm259.

27. Käll L, Krogh A, Sonnhammer EL: **A Combined Transmembrane Topology and Signal Peptide Prediction Method**. *Journal of Molecular Biology* 2004, **338**:1027-1036.

28. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N: **Expansion of the BioCyc collection of pathway/genome databases to 160 genomes**. *Nucl. Acids Res.* 2005, **33**:6083-6089.

29. Sharpton TJ, Stajich JE, Rounsley SD, Gardner MJ, Wortman JR, Jordar VS, Maiti R, Kodira CD, Neafsey DE, Zeng Q, Hung C, McMahan C, Muszewska A, Grynberg M, Mandel MA, Kellner EM, Barker BM, Galgiani JN, Orbach MJ, Kirkland TN, Cole GT, Henn MR, Birren BW, Taylor JW: **Comparative genomic analyses of the human fungal pathogens** *Coccidioides* **and their relatives**. *Genome Research* 2009, **19**:1722-1731.

30. Goffeau A, Barrell B, Bussey H, Davis R, Dujon B, Feldmann H, Galibert F, Hoheisel J, Jacq C, Johnston M, Louis E, Mewes H, Murakami Y, Philippsen P, Tettelin H, Oliver S: **Life with 6000 genes**. *Science* 1996, **274**:546-&.

31. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families**. *Nucl. Acids Res.* 2002, **30**:1575-1584.

32. Arvas M, Kivioja T, Mitchell A, Saloheimo M, Ussery D, Penttila M, Oliver S: **Comparison of protein coding gene contents of the fungal phyla Pezizomycotina and Saccharomycotina**. *BMC Genomics* 2007, **8**:325.

33. Williamson B, Tudzynski B, Tudzynski P, Kan Jalv: ***Botrytis cinerea*: the cause of grey mould disease**. *Molecular Plant Pathology* 2007, **8**:561-580.

34. Gronover CS, Schorn C, Tudzynski B: **Identification of *Botrytis cinerea* Genes Up-Regulated During Infection and Controlled by the G? Subunit BCG1 Using Suppression Subtractive Hybridization (SSH)**. *Molecular Plant-Microbe Interactions* 2004, **17**:537-546.

35. Have AT, Espino JJ, Dekkers E, Van Sluyter SC, Brito N, Kay J, González C, van Kan JA: **The *Botrytis cinerea* aspartic proteinase family**. *Fungal Genetics and Biology* 2010, **47**:53-65.

36. Schouten A, Tenberge KB, Vermeer J, Stewart J, Wagemakers L, Williamson B, Kan JALV: **Functional analysis of an extracellular catalase of *Botrytis cinerea***. *Molecular Plant Pathology* 2002, **3**:227-238.

37. James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J, Lumbsch HT, Rauhut A, Reeb V, Arnold AE, Amtoft A, Stajich JE, Hosaka K, Sung G, Johnson D, O'/Rourke B, Crockett M, Binder M, Curtis JM, Slot JC, Wang Z, Wilson AW, Schuszler A, Longcore JE, O'/Donnell K, Mozley-Standridge S, Porter D, Letcher PM, Powell MJ, Taylor JW, White MM, Griffith GW, Davies DR, Humber RA, Morton JB, Sugiyama J, Rossman AY, Rogers JD, Pfister DH, Hewitt D, Hansen K, Hambleton S, Shoemaker RA, Kohlmeyer J, Volkmann-Kohlmeyer B, Spotts RA, Serdani M, Crous PW, Hughes KW, Matsuura K, Langer E, Langer G, Untereiner WA, Lucking R, Budel B, Geiser DM, Aptroot A, Diederich P, Schmitt I, Schultz M, Yahr R, Hibbett DS, Lutzoni F, McLaughlin DJ, Spatafora JW, Vilgalys R: **Reconstructing the early evolution of Fungi using a six-gene phylogeny**. *Nature* 2006, **443**:818-822.

38. Pearson K: **On Lines and Planes of Closest Fit to Systems of Points in Space**. *Philosophical Magazine* 1901, **2**:559-572.

39. Giles NH, Case ME, Baum J, Geever R, Huiet L, Patel V, Tyler B: **Gene organization and regulation in the qa (quinic acid) gene cluster of *Neurospora crassa***. *Microbiol Rev* 1985, **49**:338-358.

40. Giles NH: **The Organization, Function, and Evolution of Gene Clusters in Eucaryotes**. *The American Naturalist* 1978, **112**:641-657.

41. Herron SR, Benen JAE, Scavetta RD, Visser J, Jurnak F: **Structure and function of pectic enzymes: Virulence factors of plant pathogens**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**:8762-8769.

42. Fraser JA, Heitman J: **Fungal mating-type loci**. *Current Biology* 2003, **13**:R792-R795.

43. Bisson LF, Coons DM, Kruckeberg AL, Lewis DA: **Yeast sugar transporters**. *Crit. Rev. Biochem. Mol. Biol* 1993, **28**:259-308.

44.   Cuomo C, Gueldener U, Xu J, Trail F, Turgeon B, Di Pietro A, Walton J, Ma L, Baker S, Rep M, Adam G, Antoniw J, Baldwin T, Calvo S, Chang Y, DeCaprio D, Gale L, Gnerre S, Goswami R, Hammond-Kosack K, Harris L, Hilburn K, Kennell J, Kroken S, Magnuson J, Mannhaupt G, Mauceli E, Mewes H, Mitterbauer R, Muehlbauer G, Munsterkotter M, Nelson D, O'Donnell K, Ouellet T, Qi W, Quesneville H, Roncero M, Seong K, Tetko I, Urban M, Waalwijk C, Ward T, Yao J, Birren B, Kistler H: **The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization**. *Science* 2007, **317**:1400-1402.

45.   Coutinho P, Henrissat B: **Carbohydrate-active enzymes: An integrated database approach**. In *Recent Advances in Carbohydrate Bioengineering*. Cambridge, United Kingdom: The Royal Scociety of Chemistry; 1999:3-14.

46.   Bajar A, Podila GK, Kolattukudy PE: **Identification of a Fungal Cutinase Promoter that is Inducible by a Plant Signal Via a Phosphorylated Trans-Acting Factor**. *Proceedings of the National Academy of Sciences of the United States of America* 1991, **88**:8208-8212.

47.   Skamnioti P, Furlong RF, Gurr SJ: **Evolutionary history of the ancient cutinase family in five filamentous Ascomycetes reveals differential gene duplications and losses and in *Magnaporthe grisea* shows evidence of sub- and neo-functionalization**. *NEW PHYTOLOGIST* 2008, **180**:711-721.

48.   Pazzagli L, Pantera B, Carresi L, Zoppi C, Pertinhez T, Spisni A, Tegli S, Scala A, Cappugi G: **Cerato-platanin, the first member of a new fungal protein family**. *Cell Biochemistry and Biophysics* 2006, **44**:512-521.

49.   Gerhold DL, Pettinger AJ, Hadwiger LA: **Characterization of a plant-stimulated nuclease from *Fusarium solani***. *Physiological and Molecular Plant Pathology* 1993, **43**:33-46.

50.   Cook RJ: **Making Greater Use of Introduced Microorganisms for Biological Control of Plant Pathogens**. *Annu. Rev. Phytopathol.* 1993, **31**:53-80.

51.   Zhang N, Castlebury LA, Miller AN, Huhndorf SM, Schoch CL, Seifert KA, Rossman AY, Rogers JD, Kohlmeyer J, Volkmann-Kohlmeyer B, Sung G: **An overview of the systematics of the Sordariomycetes based on a four-gene phylogeny**. *Mycologia* 2006, **98**:1076-1087.

52.   Suzuki C, Ando Y, Machida S: **Interaction of SMKT, a killer toxin produced by *Pichia farinosa,* with the yeast cell membranes**. *Yeast* 2001, **18**:1471-1478.

53.   Martinez D, Challacombe J, Morgenstern I, Hibbett D, Schmoll M, Kubicek CP, Ferreira P, Ruiz-Duenas FJ, Martinez AT, Kersten P, Hammel KE, Vanden Wymelenberg A, Gaskell J, Lindquist E, Sabat G, Splinter BonDurant S, Larrondo LF, Canessa P, Vicuna R, Yadav J, Doddapaneni H, Subramanian V, Pisabarro AG, Lavín JL, Oguiza JA, Master E, Henrissat B, Coutinho PM, Harris P, Magnuson JK, Baker SE, Bruno K, Kenealy W, Hoegger PJ, Kües U, Ramaiya P, Lucas S, Salamov A, Shapiro H, Tu H, Chee CL, Misra M, Xie G, Teter S, Yaver D, James T, Mokrejs M, Pospisek M, Grigoriev IV, Brettin T, Rokhsar D, Berka R, Cullen D: **Genome,**

transcriptome, and secretome analysis of wood decay fungus *Postia placenta* supports unique mechanisms of lignocellulose conversion. *Proceedings of the National Academy of Sciences* 2009, **106**:1954-1959.

54. Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivarès CP: **Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi***. *Nature* 2001, **414**:450-453.

55. Hurst CJ, Knudsen GR, McInerney MJ, Stetzenbach LD, Walter MV, Knudsen GR: *Manual of Environmental Microbiology*. 3rd edition. ASM Press; 1997.

56. Lynch M, Conery JS: **The Evolutionary Fate and Consequences of Duplicate Genes**. *Science* 2000, **290**:1151-1155.

57. Galagan J, Selker E: **RIP: the evolutionary cost of genome defense**. *Trends in Genetics* 2004, **20**:417-423.

58. Martin F, Aerts A, Ahren D, Brun A, Danchin EGJ, Duchaussoy F, Gibon J, Kohler A, Lindquist E, Pereda V, Salamov A, Shapiro HJ, Wuyts J, Blaudez D, Buee M, Brokstein P, Canback B, Cohen D, Courty PE, Coutinho PM, Delaruelle C, Detter JC, Deveau A, DiFazio S, Duplessis S, Fraissinet-Tachet L, Lucic E, Frey-Klett P, Fourrey C, Feussner I, Gay G, Grimwood J, Hoegger PJ, Jain P, Kilaru S, Labbe J, Lin YC, Legue V, Le Tacon F, Marmeisse R, Melayah D, Montanini B, Muratet M, Nehls U, Niculita-Hirzel H, Secq MPO, Peter M, Quesneville H, Rajashekar B, Reich M, Rouhier N, Schmutz J, Yin T, Chalot M, Henrissat B, Kues U, Lucas S, Van de Peer Y, Podila GK, Polle A, Pukkila PJ, Richardson PM, Rouze P, Sanders IR, Stajich JE, Tunlid A, Tuskan G, Grigoriev IV: **The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis**. *Nature* 2008, **452**:88-92.

59. Lorenz N, Olsovska J, Sulc M, Tudzynski P: **The alkaloid cluster gene ccsA of the ergot fungus *Claviceps purpurea* encodes the chanoclavine-I-synthase, an FAD-containing oxidoreductase mediating the transformation of N-methyl-dimethylallyltryptophan to chanoclavine-I**. *Appl. Environ. Microbiol.* 2010:AEM.00737-09.

60. Vanden Wymelenberg A, Sabat G, Mozuch M, Kersten P, Cullen D, Blanchette R: **Structure, organization, and transcriptional regulation of a family of copper radical oxidase genes in the lignin-degrading basidiomycete *Phanerochaete chrysosporium***. *Appl. Environ. Microbiol.* 2006, **72**:4871-4877.

61. Lawrence JG, Roth JR: **Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters**. *Genetics* 1996, **143**:1843-1860.

62. Martinez DA, Nelson MA: **The Next Generation Becomes the Now Generation**. *PLoS Genet* 2010, **6**:e1000906.

63. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995, **57**:289-300.

64. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G: **GO: TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes**. *Bioinformatics* 2004, **20**:3710-3715.

65. Altschul SF, Boguski MS, Gish W, Wootton JC: **Issues in searching molecular sequence databases**. *Nat Genet* 1994, **6**:119-129.

66. Holland RCG, Down TA, Pocock M, Prlic A, Huen D, James K, Foisy S, Drager A, Yates A, Heuer M, Schreiber MJ: **BioJava: an open-source framework for bioinformatics**. *Bioinformatics* 2008.

CONCLUSIONS AND FUTURE DIRECTIONS

In this dissertation, we have shown that large portions of all fungal genomes have a non-random organization of genes. By examining different types of gene clusters in many genomes, we have established this as an important phenomenon in genome evolution that should play a key role in future work.

In the first chapter, we established that in at least one genome (*Trichoderma reesei*), key genes producing biomass degrading proteins have a nonrandom organization, and form clusters in many areas of the genome. Another important parameter we established is that the architecture of the genome is reorganizing to create optimal configurations in these areas, changing drastically since divergence from several relatives. However, many unknowns about gene cluster formation remain.

In the second chapter, we were able to catch a glimpse of how genomes evolve between haplotypes of the same organism. While large scale changes were not detected, changes in the frequency of nucleotide substitutions with regard to the location of the gene were apparent. In this study, we also showed that genes thought to be involved in the peculiar method of cellulose degradation in *Postia placenta* are likely clustered in non-syntenic areas of the genome when compared to other biomass degrading relatives such as *Phanerochaete chrysosporium*. This is similar to what was found in *T. reesei*. This organization suggests that the biomass degrading genes in *P. placenta* expanded after the common ancestor of *P. placenta* and *P. chrysosporium* diverged, or else migrated to different areas than in *P. chrysosporium*, which contains similar genes thought to be involved in biomass degradation that are also found in clusters [1].

In our third and final study, we showed that in all sequenced fungal genomes contain gene clusters. To execute this study across an unprecedented number of genomes, we created a robust and flexible algorithm that completes in trivial time. By performing this study in many genomes, we were able to compare clustering across the fungal kingdom and discover trends concerning the clusters on a global scale. This data set will serve as ample substrate for our future investigations for years to come. We will make all data available to assist other researchers in their investigations and comparative analyses. Also, we will be able to quickly deploy the pipeline for any number of genomes that are fully sequenced, be they from the fungal kingdom or from any other branch of life.

While these studies have provided information about existing clusters and their characteristics, much of how clusters form remains unknown. One hypothesis we were not able to test is the horizontal gene transfer formation scenario (HGT) [2]. While HGT has long been implicated in gene clustering, it remains controversial, as it is difficult to demonstrate conclusively. We have shown that while some duplication is involved in cluster formation, it is limited to a few cases. Instead, most formation that we detected can be attributed to gene movements. We cannot, however, easily distinguish between the movement of genes within a genome and the transfer of genes into the genome from a foreign source. In our study, we assumed that the simplest and thus most likely method of formation was by gene movements; however, it is probable that HGT was the cause of a small amount of clustering [3, 4]. Work in our laboratory continues to provide insight into HGT clustering on a kingdom wide scale.

One aspect of gene clustering that has been more difficult to investigate is the maintenance of clusters. Once clusters form, do they change, or do they remain the same? Much of our

data in Chapter 3 supports independent formation and suggests there is very little maintenance; while some clusters appear to be conserved, in several cases different genes are involved. The low number of MCL groups that appear across all species suggests that they are not maintained at a high level. This conclusion, however, is tainted by paucity of data, as we were not able to include many of the closely related relatives that were included in the broader study. In future work, we will recalculate all genomes and thus be able to provide a clearer answer. Additionally, we will compare the characteristics of the clustered versus the non-clustered regions.

Why do genes cluster? A popular hypothesis, which is integrated into the HGT hypothesis, is that coregulation drives cluster formation and is also responsible for maintenance of gene clustering. This hypothesis could be tested through gene expression studies. One downside is that these studies can be expensive and difficult to perform. Another point to consider is the condition(s) under which a particular cluster might be induced. Thus while several microarray or RNASeq experiments may be conducted to test the clusters for coexpression, it will be difficult to demonstrate that any cluster is false, since perhaps the right growth condition has simply not been found. A definitive null model remains difficult to define, and only true positives constitute reliable evidence. At least one study has suggested that a clustered region in *P. chrysosporium* is not coordinately regulated [5]. For reasons that we discuss below, this still may not be enough evidence to completely rule out the possibility that involvement in a related process is forcing the genes to colocalize in the genome.

The gene clusters may reflect a broader and unknown level of nuclear coordination that controls the three dimensional assemblage of chromatin. A known nuclear pore contains a zinc-finger (i.e. DNA binding) domain [6]. This suggests that the nuclear pore might be able

166

to "pin" the chromatin to the nuclear membrane, exposing stretches of DNA to a common transcriptional machinery. Thus highly correlated cotranscription may not be necessary to keep clusters together; instead, a gene or set of genes may simply need to be near the exposed region for subsequent easier access. Indeed, hours to days separate the increase in different monoxygenase transcript abundance [5]. In this scenario, it is also possible that separation of genes by some distance may be advantageous, as shorter stretches of DNA emerge from the spaghetti of chromatin inside the nucleus. This model supports the separation in clusters seen in our study and others [7-9].

The identification of gene clusters in fungal genomes has impact on many areas of research. These clustered regions highlight areas of the genome that are of immense importance to the particular fungus, in that they often contain lineage specific genes that are key to species specific abilities, as well as being sorted into an optimal configuration. By analyzing genomes spanning the entire fungal kingdom, we have identified genomic regions that are critical to organisms that are agricultural pathogens, biomass decomposers and animal pathogens. Thus our study impacts research in medicine, biofuels, biotechnology and global food production. Not only does this body of research add value to applied science research, but also to research in basic cell biology. The same key regions of the genome could add to the three dimensional model of chromatin configuration within the nucleus, as clustered regions may be pulled toward the nuclear membrane.

As new genomes become available, we will continue to add fungal genomes as well as the genomes of other organisms. By developing the cornerstone of genomic architecture research, we have laid the foundations in this field for new investigators to follow.

References

1.   Martinez D, Larrondo L, Putnam N, Gelpke M, Huang K, Chapman J, Helfenbein K, Ramaiya P, Detter J, Larimer F, Coutinho P, Henrissat B, Berka R, Cullen D, Rokhsar D: **Genome sequence of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78**. *Nature Biotechnology* 2004, **22**:695-700.

2.   Lawrence JG, Roth JR: **Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters**. *Genetics* 1996, **143**:1843-1860.

3.   Mallet L, Becq J, Deschavanne P: **Whole genome evaluation of horizontal transfers in the pathogenic fungus Aspergillus fumigatus**. *BMC Genomics* 2010, **11**:171.

4.   Khaldi N, Wolfe KH: **Elusive Origins of the Extra Genes in Aspergillus oryzae**. *PLoS ONE* 2008, **3**:e3036.

5.   Vanden Wymelenberg A, Sabat G, Mozuch M, Kersten P, Cullen D, Blanchette R: **Structure, organization, and transcriptional regulation of a family of copper radical oxidase genes in the lignin-degrading basidiomycete Phanerochaete chrysosporium**. *Appl. Environ. Microbiol.* 2006, **72**:4871-4877.

6.   Sukegawa J, Blobel G: **A nuclear pore complex protein that contains zinc finger motifs, binds DNA, and faces the nucleoplasm**. *Cell* 1993, **72**:29-38.

7.   Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, Chapman J, Chertkov O, Coutinho PM, Cullen D, Danchin EGJ, Grigoriev IV, Harris P, Jackson M, Kubicek CP, Han CS, Ho I, Larrondo LF, de Leon AL, Magnuson JK, Merino S, Misra M, Nelson B, Putnam N, Robbertse B, Salamov AA, Schmoll M, Terry A, Thayer N, Westerholm-Parvinen A, Schoch CL, Yao J, Barbote R, Nelson MA, Detter C, Bruce D, Kuske CR, Xie G, Richardson P, Rokhsar DS, Lucas SM, Rubin EM, Dunn-Coleman N, Ward M, Brettin TS: **Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina)**. *Nat Biotech* 2008, **26**:553-560.

8.   Lee JM, Sonnhammer EL: **Genomic Gene Clustering Analysis of Pathways in Eukaryotes**. *Genome Res.* 2003, **13**:875-882.

9.   Yi G, Sze S, Thon MR: **Identifying clusters of functionally related genes in genomes**. *Bioinformatics* 2007, **23**:1053-1060.