LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Evolutionary Genomics of Filamentous Fungi

by

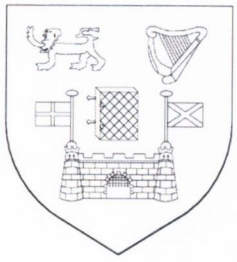
Nora Khaldi
DESS

A thesis submitted to
The University of Dublin
for the degree of

Doctor of Philosophy

TRINITY COLLEGE
0 6 NOV 2008
LIBRARY DUBLIN

Department of Genetics
Trinity College
University of Dublin



April 2008

THESIS
8605

Declaration

This thesis has not been submitted as an exercise for a degree at any other University. Except where otherwise stated, the work described herein has been carried out by the author alone. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College. The copyright belongs jointly to the University of Dublin and Nora Khaldi.

Signature of Author..... .....

Nora Khaldi

30th April 2008

Summary

Filamentous fungi are important producers of secondary metabolites (SMs), together with plants and bacteria. While some SMs are toxic to humans, animals or plants, many others are beneficial. The discovery that the genes for each SM biosynthesis pathway tend to be physically clustered into a single locus in the genomes of filamentous fungi raises many evolutionary questions such as why, where and how do these clusters emerge? In this thesis I investigate some evolutionary questions about the origins of SM gene clusters, using the fumonisin cluster of *Fusarium verticillioides* and the Ace1 cluster of *Magnaporthe grisea* as models. I show that the fumonisin cluster was formed *de novo* in a recent ancestor of *F. verticillioides*; other filamentous fungi including the closely related *F. oxysporum* have orthologs of many of the genes, but these genes are not physically clustered. A different mechanism was found to underlie the formation of the Ace1 cluster: in this case, both duplication of a pre-existing cluster and horizontal gene transfer were involved.

To facilitate the identification and study of SM gene clusters across large numbers of sequenced fungal genomes, I developed the software package SMURF (Secondary Metabolite Unknown Regions Finder). This program identifies putative SM gene clusters based on the presence of one of the four main classes of SM biosynthesis backbone genes, closely flanked by other genes whose products contain one of 27 SM-associated domains that we identified by examination of known fungal SM gene clusters. SMURF was found to be able to recover known or manually annotated SM gene clusters with high sensitivity, and even in the well-studied genome of *Aspergillus fumigatus* SMURF identified several putative clusters that were not detected previously.

Horizontal gene transfer, implicated in the origin of the Ace1 cluster, is a phenomenon best known in prokaryotes. Very few individual transfers of single genes have been identified in multicellular eukaryotes. I investigated horizontal transfer as a possible cause of the increased number of genes in the genome of

Aspergillus oryzae as compared to the closely related species *A. fumigatus* and *A. nidulans*. I found that at least twelve genes were horizontally transferred into *A. oryzae* from the Sordariomycetes group and estimate that this is only the tip of the iceberg.

The two final chapters of this thesis report work that I did as part of large international genome-sequencing consortia. In one, I compare the genome sequences of *A. flavus* and *A. oryzae* to identify genes that are evolving rapidly in one genome relative to the other. I present evidence for widespread relaxation of selective constraints on *A. oryzae* genes, consistent with it being a domesticated species. In the other chapter, I contributed to the analysis of inter-strain and inter-species genomic differences among *A. fumigatus*, *A. clavatus* and *Neosartorya fischeri*.

Acknowledgements

So many people have provided me with support, insight and inspiration during the last three and half years. I would first like to thank my supervisor Prof. Ken Wolfe, who has given me his expert guidance, experience and patience from my first days in Trinity to my final submission. The Wolfe lab in Trinity has been more than a place of work for me, it has been my home from home with many dear friends and colleagues. In particular I would like to thank Kevin, Gavin and Meg for their help both with science and life. I would also like to express my gratitude to current and past Wolfe Lab family members, Brian, Devin, Marie, Nadia, Jonathan, Jeff, Carolin, Simon, John, Matt and Sean. I would also like to thank my international collaborators including Bill and Natalie from JCVI, Gary and his lab in North Carolina, and Jerome in Lyon. I had a wonderful and beneficial time in all their labs.

I would also like to thank my friends, Efi, Brian and of course my chocolate provider Lillian.

A big thank you to a very special person to me, Ciaran, for his support, love, fun and cooking skills!!

Last but by no means least I would like to thank the Khaldi clan in Marseille, Maman, Papa and Joe. Although we are separated by a long distance I have felt you by my side throughout. Your visits, calls and words of encouragement have been a constant support.

For my parents

Table of Contents

CHAPTER 1 INTRODUCTION **1**

1.1 Overview 1

1.2 Filamentous fungi 1

1.3 Secondary metabolism 2

1.3.1 Structural diversity of SMs 4

1.3.2 Backbone enzymes of secondary metabolism 5

1.3.2.1 Polyketides 5

1.3.2.2 Non-Ribosomal Peptide Synthetases 10

1.3.2.3 Hybrid PKS-NRPS enzymes 12

1.3.2.4 Terpenoids 12

1.3.2.5 Alkaloids 14

1.3.3 Genes in secondary metabolite pathways 16

1.4 Physical clustering of genes on chromosomes 17

1.4.1 Background 17

1.4.2 Clustering of secondary metabolite clusters in filamentous fungi 19

1.4.3 Why do fungi synthesize SMs? 21

1.4.4 Why do SM genes cluster? Is there a unifying explanation for the clustering of SM biosynthesis genes in filamentous fungi? 23

1.4.4.1 The inheritance model 23

1.4.4.2 Selectionist theory 26

1.4.4.3 Coadapted gene complexes 31

1.4.4.4 Conclusion 31

CHAPTER 2 EVOLUTIONARY ORIGINS OF THE FUMONISIN SECONDARY METABOLITE GENE CLUSTER IN *FUSARIUM VERTICILLIOIDES* **33**

2.1 Abstract 33

2.2 Introduction 34

2.3 Methods 35

2.4 Results 36

2.4.1 Recent assembly of the FUM gene cluster 36

2.4.2 Accelerated sequence evolution of the fumonisin cluster genes 40
2.5 Discussion 42

CHAPTER 3 SMURF: A WEB-BASED TOOL FOR IDENTIFICATION OF FUNGAL SECONDARY METABOLITE BIOSYNTHESIS GENE CLUSTERS 44
3.1 Abstract 44
3.2 Introduction 45
3.3 Methods 48
3.3.1 Detecting backbone enzymes. 48
3.3.2 Detecting SM clusters. 49
3.4 Results 50
3.4.1 Overview of SMURF 50
3.4.2 Training set 51
3.4.3 Application to *Aspergillus fumigatus* 53
3.4.4 Uneven taxonomic distribution of backbone enzymes 56
3.4.5 Species specificity of SM clusters 59
3.5 Discussion 61

CHAPTER 4 EVIDENCE FOR HORIZONTAL TRANSFER OF A SECONDARY METABOLITE GENE CLUSTER BETWEEN FUNGI 65
4.1 Related publication 65
4.2 Abstract 65
4.3 Introduction 66
4.4 Methods 68
4.5 Results 69
4.5.1 Identification of homologous ACE1 clusters in other filamentous fungi 69
4.5.2 Phylogenetic analysis of the ACE1 cluster in filamentous fungi 71
4.5.3 Identifying the direction of gene transfer 75
4.6 Discussion 76
4.6.1 The ACE1 cluster is specific to few fungal species 76
4.6.2 ACE1 cluster evolution in Sordariomycetes involved several duplication events 78

4.6.3 HGT of a fungal secondary metabolism gene cluster 80

CHAPTER 5 ELUSIVE ORIGINS OF THE EXTRA GENES IN *ASPERGILLUS ORYZAE* 83

5.1 Abstract 83

5.2 Introduction 84

5.3 Methods 86

5.3.1 Data 86

5.3.2 Identification of Sets S_{AO}, S_{AN} and S_{AFU} 86

5.3.3 Testing for WGD 87

5.3.4 Testing for HGT 88

5.4 Results 88

5.4.1 An excess of topology S trees 88

5.4.2 Topology S trees are not artifacts of rate acceleration after gene duplication 89

5.4.3 No evidence for ancient WGD in *Aspergillus* 91

5.4.4 Can the excess of Topology S genes be explained by Horizontal Gene Transfer? 93

5.4.5 Functions and physical clustering of AO2 genes with Topology A within the *A. oryzae* genome 98

5.5 Discussion 100

5.6 Acknowledgments 102

CHAPTER 6 EFFECTS OF DOMESTICATION ON GENOME EVOLUTION IN *ASPERGILLUS FLAVUS* AND *A. ORYZAE* 103

6.1 Abstract 103

6.2 Introduction 104

6.3 Methods 105

6.4 Results and Discussion 106

CHAPTER 7 GENOMIC ISLANDS IN THE PATHOGENIC FILAMENTOUS FUNGUS *ASPERGILLUS FUMIGATUS* 111

7.1 Publication 111

7.2 Abstract 112

7.3 Introduction 112
7.4 Methods 115
7.4.1 Fungal Isolates 115
7.4.2 Accession Numbers 115
7.4.3 Whole Genome Sequencing 115
7.4.4 Sequence identity at the nucleotide level 115
7.4.5 Gene structure annotation 116
7.4.6 Repetitive elements 116
7.4.7 A. fumigatus annotation improvements 116
7.4.8 Functional annotation 117
7.4.9 Ortholog identification 117
7.4.10 Synteny analysis 118
7.4.11 Statistical analysis 118
7.4.12 Selective constraints 118
7.4.13 Phylogenetic analyses 119
7.4.14 Southern blot analysis 120
7.4.15 Colony radial growth rate measurement 120
7.5 Results and discussion 120
7.5.1 A. fumigatus Af293 vs. *A. fumigatus* A1163 121
7.5.2 Highly Variable Loci in A. fumigatus 123
7.5.3 A. fumigatus vs. *N. fischeri* vs. *A. clavatus* 125
7.5.4 Core and Lineage-Specific Genes 127
7.5.4.1 Features of Core and Lineage-Specific Genes 127
7.5.4.2 Biological Roles and Chromosomal Location of LS Genes 131
7.5.4.3 Lineage Specific Genomic Islands 134
7.5.4.4 Evolutionary Origins of Lineage-Specific Genes 135
7.5.5 Conservation of Virulence-, Allergy-, and Sex-Associated Genes 138
7.6 Conclusion 140

CHAPTER 8 141

DISCUSSION 141

APPENDIX A 143

REFERENCES **145**

Chapter 1

Introduction

1.1 Overview

In this chapter I provide general information on filamentous fungi, and I review the literature on what is known about secondary metabolism and the genes implicated in secondary metabolite (SM) biosynthesis in filamentous fungi. I discuss and provide a background on the evolution of clustering of genes in general, and clustering of SM genes in particular. I also review the literature of horizontal gene transfer of nuclear genes between eukaryotes.

1.2 Filamentous fungi

Seven phyla constitute the fungal world (Hopwood 1997; Mootz et al. 2002; Keller et al. 2005; Hibbett et al. 2007). All fungi share important morphological and biochemical features, such as walls typically containing chitin. Five of the seven phyla are relatively little known: Chytridiomycota, commonly known as chytrids; Blastocladiomycota; Neocallimastigomycota; Zygomycota, which contains the taxa Zygomycetes and Trichomycetes; and Glomeromycota. The two better-known phyla are Basidiomycota and Ascomycota. Basidiomycota, commonly known as basidiomycetes, includes all common mushrooms and rusts, the maize pathogen *Ustilago maydis* and the opportunistic human pathogen *Cryptococcus neoformans*. Ascomycota, commonly known as ascomycetes, are fungi that form meiotic spores called ascospores, which are enclosed in a special

sac-like structure called an ascus. This division includes morels, a few mushrooms and truffles, single-celled yeasts (*e.g.*, *Saccharomyces*, *Kluyveromyces*, *Pichia* and *Candida*), and many filamentous fungi which are my focus in this thesis. I will also use some basidiomycete species for comparison reasons.

Filamentous fungi live as saprotrophs, parasites, and mutualistic symbionts. Prominent and important genera of filamentous ascomycetes include *Aspergillus*, *Fusarium*, *Penicillium*, *Claviceps* and *Magnaporthe*. These fungi produce filaments termed hyphae. Fungal hyphae grow and branch to produce a network of filaments that constitute the mycelium. The mycelium enlarges by extension of single hyphae, which show polar growth, meaning they grow only at their extreme tips. This apical growth is in contrast to the intercalary growth of most other filamentous organisms. Expansion of the mycelium is continuous if the hyphae can keep on extending on the medium they are residing in. However, changes do take place as the mycelium ages and as that part of the food source on which it is growing is no longer able to provide sufficient nutrients (Fawcett 1925; Ryan et al. 1943). Gottlieb however, argues against the continuous growth of the hyphae even in the presence of sufficient nutrients (Gottlieb 1971).

1.3 Secondary metabolism

Fungi can reproduce by both asexual and sexual means. Reproduction is invariably connected to the production of spores, produced at specialized structures. Spores are fully equipped to start a new colony independent of the parent mycelium, and usually some distance from it. Fungal spores vary enormously in shape, size and other special properties, linked to their numerous roles in dispersal or survival. Many ascomycete species have only been observed undergoing asexual reproduction (these are called anamorphic species), but molecular data has often been able to identify their closest teleomorphs in the Ascomycota. Because the products of meiosis are retained within the sac-like ascus, several ascomycetes (most prominently, *Neurospora crassa* and *Aspergillus nidulans*) have played a central role in elucidating the principles of genetics and heredity.

Secondary metabolites (SMs), also known as natural products, are chemicals produced by all organisms, ranging from prokaryotes to eukaryotes. The production of SMs is particularly enriched in bacteria, plants and fungi. Unlike primary metabolites that are often essential for the species survival and reproduction and are implicated in many pathways such as energy and protein production, SM are, as their name suggests, secondary to the organism's survival. Human awareness of SMs dates back to the 19th century but it is only in the last few decades that secondary metabolism has attracted the interest of microbiologists (Bentley and Bennett 1999).

SMs are produced by a limited number of precursors derived from primary metabolism. The classification of SMs thus depends on the precursor utilized (Keller et al. 2005). Four main classes of SMs exist:

- Polyketides are the outcome of the condensation of acyl-coenzyme A. Polyketides are made by polyketide synthase (PKS) enzymes.
- Non-ribosomal-peptides are peptides synthesized from amino acid precursors independently of the ribosomal translation mechanism. They are made by non-ribosomal peptide synthase (NRPS) enzymes.
- Terpenes are composed of isoprene units, dimethylallyl diphosphate, produced by the mevalonate pathway. The key enzymes in terpene synthesis are the terpene cyclases.
- Alkaloids are also composed of dimethylallyl diphosphate units, but can be made out of tryptophan (by the enzyme dimethylallyl tryptophan synthetase, prenyl transferase), and more rarely other amino acids.

Polyketides and non-ribosomal peptides are the most common classes of SM in fungi.

1.3.1 Structural diversity of SMs

Although each SM class follows a single process for its synthesis, a diverse range of structural differences exists within each class. This diversity affects the final molecule and is mainly due to the number of precursors utilized, the different modifications during the elongation process, and the choice of extender units (for example, sugars) issued from other biosynthetic pathways.

It has been shown that for certain known fungal SMs, the species also produces analogues, known as chemotypes, at lower concentrations. For example, *Fusarium verticillioides* produces the polyketide fumonisin B (FB) that can have different forms depending on the presence or absence of hydroxyl groups on carbons 5 and 10 (Bojja et al. 2004). FB1 is the most abundant fumonisin produced by *F. verticillioides*, but this filamentous fungus also produces FB2, FB3 and FB4 in lower amounts. Chemotype diversity can reach very high levels, most strikingly in entomopathogenic *Metarhizium anisopliae* in which 27 chemotypes of destruxin were characterized (Pedras et al. 2002).

Due to the important range of chemical structures of SMs, their biological activities are also very diverse. One can classify SMs into different groups depending on many factors such as their impact on human, animal, or plant health.

Fungal SMs beneficial to humans and animals include many antibiotics, many of which are non-ribosomal peptides (penicillin is a good example). Other beneficial SMs include immunosuppressors such as cyclosporin produced by *Trichoderma polysporum*, which is utilized in renal transplantation (Ponticelli 2005), and anti cholesterol molecules from the lovastatin family, produced by *A. terreus* (Alberts et al. 1980). On the other hand many SMs are detrimental to human and animal health. Many narcotics are SMs. For example, LSD – the most hallucinogenic molecule known – is synthesized from Lysergic acid which is an alkaloid produced by the fungus *Claviceps purpurea* (Schardl et al. 2006). Most of the

toxins implicated in food toxicity are also SMs, the best known being aflatoxin produced by some *Aspergillus* species.

Some SMs are classified as phytotoxins. These are molecules produced by a phytopathogenic fungus, that are implicated in plant infection and destruction, and can reach epidemiological levels. T-toxin is a good example of such an epidemic that devastated American maize cultures in 1970.

We can also classify SMs on their indirect effect on humans and animals when their producer fungi infect plants. Many metabolites are zootoxic. As a consequence, the presence of these SMs in plant material such as grains affects human and animal health. Aflatoxin B1 is the most cancerous natural product known to date (Squire 1981). Even small sampling of cereal plants has shown that trichothecenes deoxynivalenol and nivalenol, zearalenone and fumonisins are produced at dangerous levels on wheat and corn, cereals that are utilized in human and animal foodstuffs (Placinta et al. 1999).

The SMs that have been studied intensively are those such as cyclosporin A or lovastatin that have had an enormous impact on human or animal health. In stark contrast, many other metabolites receive very little attention because their impact on human society is much more insignificant; examples include patulin, fusicoccin, and cercosporin.

1.3.2 Backbone enzymes of secondary metabolism

1.3.2.1 Polyketides

1.3.2.1.1 Structure and classification of polyketide synthases (Figure 1.1)

Polyketide synthase (PKS) enzymes possess different domains implicated in the different reactions that are essential in the synthesis of polyketides. The KS domain (beta-ketoacyl CoA synthesis), AT domain (acyl transferase), and ACP domain (acyl carrier protein) all catalyze the elongation of the carbon chain. The

KR domain (beta-keto reductase), DH domain (dehydratase) and ER domain (enoyl reductase) are responsible for the formation of the alkyls.
Polyketides are very similar to fatty acids and both have the same precursor, usually acetyl coenzyme A (acetyl CoA) and malonyl CoA. Fungal PKSs are closely related to fatty acid synthases (FASs). Both are multifunctional enzymes, and both share the minimal structure of domains, this is: KS, AT, and ACP (Keller et al. 2005) (Figure 1.1) Whereas the KR, DH, and ER domains are present in all FASs, some or all are absent in various PKSs. KR, DH, and ER domains catalyze, reduction of a keto to a hydroxyl group, dehydration of the hydroxyl to an enoyl group, and reduction of the enoyl to an alkyl group, respectively (Figure 1.1) Several factors contribute to the diversity of polyketides: the number of iterations, the number of reduction reactions, the addition of extender units, such as sugars, and, for aromatic polyketides, the cyclizations of the nascent polyketide (Hopwood and Sherman 1990).

We distinguish different types of PKS depending on their structure (Du et al. 2001; Shen 2003; Watanabe et al. 2007):

- Type I polyketide synthase are multifunctional enzymes in which each catalytic domain is utilized iteratively. The majority of fungal PKS are of this type. In actinomycete bacteria like *Streptomyces* we find modular type I PKSs producing aromatic polyketides. Each module possesses the structure of a type I PKS and is only utilized once during the synthesis of a metabolite.
- Type II polyketide synthase are enzymatic complexes in which each domain is coded by a different gene. Each gene is used iteratively to produce a polyketide. This the most common type of PKS in bacteria.
- Type III polyketide synthase are atypical enzymes that correspond to a KS domain that functions iteratively to synthesize aromatic polyketides. A well known example from plants is chalcone synthase, but this type of PKS was only recently identified in fungi (Seshime et al. 2005). These enzymes do not possess ACP domains. They utilize Coenzyme A thioester as precursors.

Besides the above classification, fungal type I PKSs can also be classified phylogenetically based on their domain content. Kroken and co-workers showed that we can phylogenetically distinguish different clades of PKS with different structural organization that can give rise to reduced, partially reduced or fully oxidized PKS (Kroken et al. 2003). It is the KR, DH, and ER domains that catalyze reduction of a keto to a hydroxyl group, dehydration of the hydroxyl to an enoyl group, and reduction of the enoyl to an alkyl group, respectively. Reducing PKS architectures include KS-AT-DH-(cMeT)-ER-KR-ACP (*e.g.*, fumonisin PKS), or KS-AT-DH-ER-KR-ACP-(ACP) (*e.g.*, T-toxin biosynthesis PKS), or KS-AT-DH-(cMeT)-KR-ACP-(C)-(A)-(PCP) (*e.g.*, lovastatin PKS). Non reducing PKS architectures include KS-AT-ACP-(ACP)-(TE) or KS-AT-ACP-

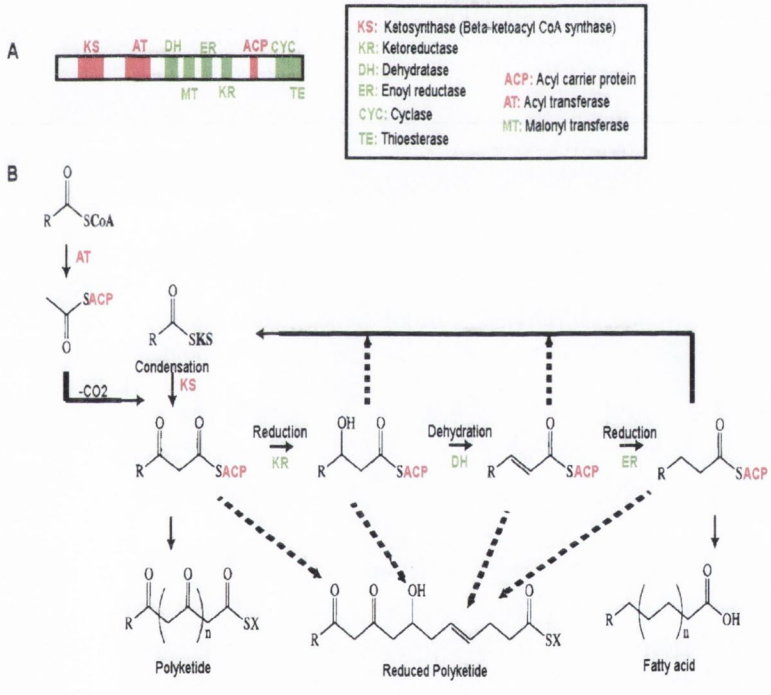


Figure 1.1 This figure is reproduced from (Hopwood 1997) and (Yu et al. 2004; Keller et al. 2005). It shows the organization and enzymatic reactions catalyzed by the iterative type I PKS domains. A. Red domains are the minimal structure required by a PKS to produce a polyketide. Optional domains are in green and are implicated in the modification and release of the polyketide. The modules are used iteratively, they are only used once in modular PKSs. KS: β-ketoacyl CoA synthase; AT: acyl transferase; ACP: acyl carrier protein. DH: dehydratase; cMeT : C-methylase; ER : enoyl reductase; KR : β-keto reductase; TE : thioesterase. B. The AT domain initiates the polyketide synthesis by loading acetyl and malonyl coenzyme A (CoA) as thioesters on to the 4'-phosphopantotheine of an acyl carrier (ACP) domain. Condensation then occurs with another thioester intermediate bound to the ketoacyl CoA synthase (KS) domain, and decarboxylation of the ACP-bound intermediate occurs. The ketoreductase (KR) domain then reduces the resulting β-ketothioester, followed by a dehydration by the dehydratase (DH) domain. If an enoyl reductase (ER) domain is present, an unsaturated intermediate is formed. Some PKSs contain a methyltransferase (MT) domain that methylates the α-carbon of the thioester. CYC, cyclase; TE, thioesterase.

(ACP)-cMet-(TE). Here, the acronyms cMeT and TE refer to domains for methylation and the release of the polyketide, respectively.

1.3.2.1.2 Enzymatic reactions catalyzed by Type I PKS

1.3.2.1.2.1 ACP domain activation

The ACP domain of the PKS need to be first activated. The activation process consists of the transfer of a cofactor 4'-phosphopantetheine (4'-PP) onto a conserved serine of the ACP domain (Mootz et al. 2002) (Figure 1.1) The transfer is catalyzed by a 4'-PP transferase (PPTase). Many PPTase genes exist in bacteria versus only one PPTase gene implicated in secondary metabolism in *A. nidulans* (*npgA*), which also has a homolog in *A. fumigatus* (Keszenman-Pereyra et al. 2003; Oberegger et al. 2003; Marquez-Fernandez et al. 2007). The PPTases are required in fungi for the production of penicillin, pigments, siderophore and lysins. The fact that there are only one in fungi compared to a few in bacteria suggests that these genes activate a much larger spectrum of ACP and PCP domains (Marquez-Fernandez et al. 2007).

1.3.2.1.2.2 Acyl condensation

The three minimal domains of a PKS – KS, AT and ACP – are implicated in the condensation of the thioester. The AT domain catalyses the transfer of the acyl CoA on the prosthetic 4'-PP of the ACP domain (Figure 1.1) The carbanion that is then formed reacts with the acyl that is carried by a conserved cysteine residue of the KS domain (Hopwood 1997; Schwarzer and Marahiel 2001) (Figure 1.1)

1.3.2.1.2.3 Beta-ketone modification

The resulting β-ketothioester goes through many modifications. The KR domain catalyses the reduction of the ketone to a hydroxyl using a NAD(P)H molecule. The resulting molecule can then undergo a dehydration catalyzed by the DH domain. Finally, the enoyl reductase (ER) domain can reduce the resulting enoyl to give rise to an unsaturated intermediate.

1.3.2.1.2.4 Release of the polyketide from the polyketide synthases

The C-termini of some PKSs contain a thioesterase (TE) domain that is implicated in the release of the polyketide. This will result in the liberation of a linear polyketide such as the T-toxin. Given that many PKSs do not posses this type of domain, the polyketide release might in this case be achieved by spontaneous cyclization (Song et al. 2004). There is, so far, no evidence for a catalyzed cyclization, although this cannot be ruled out. This reductive mechanism obviates the requirement for a TE domain, which is present in only some PKS modules.

1.3.2.1.2.5 Programming of the PKSs

The term "PKS program" has been introduced to describe control of the variables that determine the structure of the product of a specific PKS (Figure 1.1). These variables were described by (Hopwood 1997) as: "choice of starter unit; choice of the nature and number of the chain extender units; control of the reductive cycle on the keto group of the growing carbon chain, which in turn determines the keto, hydroxyl, enoyl, or methylene functionality at each alternate carbon atom; stereochemistry of hydroxyl and alkyl side groups; and pattern of cyclization of the nascent carbon chain". The modular program of PKSs in *Streptomyces* is well understood. The number of modules and the catalytic domains of each of them correspond exactly to the number of condensations and modifications realized to produce the polyketide. Before the first module we find an AT and ACP domain implicated in the loading of the first acyl-CoA. At the end we find a TE domain responsible for the release of the polyketide. This was confirmed by internal deletions of certain domains that resulted in the absence of predicted modifications (Hopwood 1997).

1.3.2.1.2.6 Choice of the precursor

The majority of polyketides are products of acetyl CoA and the addition of malonyl CoA molecules, but the substrates can be very diverse (Hopwood and Sherman 1990; Moore and Hertweck 2002), for instance the hexanoic acid that constitutes the start unit for the synthesis of aflatoxin (Brown et al. 1996a).

1.3.2.2 Non-Ribosomal Peptide Synthetases

1.3.2.2.1 Structure and function

Both bacterial and fungal Non-Ribosomal Peptide Synthetases (NRPSs) are modular, and have a similar organization to that of type I polyketide synthetases (Finking and Marahiel 2004). The minimal modules are a condensation domain (C), an adenylation domain (A) and a thiolation domain (T or PCP) (Figure 1.2) The C domain can often be replaced by a cyclization (Cy) domain, implicated in the cyclization and the condensation of the added amino acid (Miller and Walsh 2001). Certain modules can also contain modification domains that act prior to the amino acid condensation. These include the epimerization (E) domain that permits the incorporation of the amino acid in the D configuration, and the methylation (MeT) domain. Upstream of these modules we find the A and PCP domains permitting the loading of the first amino acid. Downstream we find the thioesterase domain (TE) implicated in the release of the peptide. Each of these domains is characterized by very conserved motifs (Marahiel et al. 1997). Similarly to the PKSs given that many NRPSs do not posses this type of domain, the non-ribosomal peptide release might in this case be achieved by spontaneous cyclization (Song et al. 2004).

1.3.2.2.2 Enzymatic reactions catalyzed by the different domain

The adenylation domain (A) domain activates the ATP hydrolysis of the amino acid to give raise to the corresponding adenylate (Figure 1.2). This unstable intermediate is then transferred to the PCP domain (Finking and Marahiel 2004) (Figure 1.2)

Once the amino acid is fixed to the PCP domain, it can undergo epimerization and methylation (Finking and Marahiel 2004). The amino acid condensation is then catalyzed by the C domain that catalyses the peptide bond formation (Stachelhaus et al. 1998). The number of C domains corresponds to the number of peptide bonds present in the linear intermediate.

The release of the peptide chain is done by hydrolysis, catalyzed by the TE domain that is found in all NRPS (Marahiel et al. 1997) (Figure 1.2).

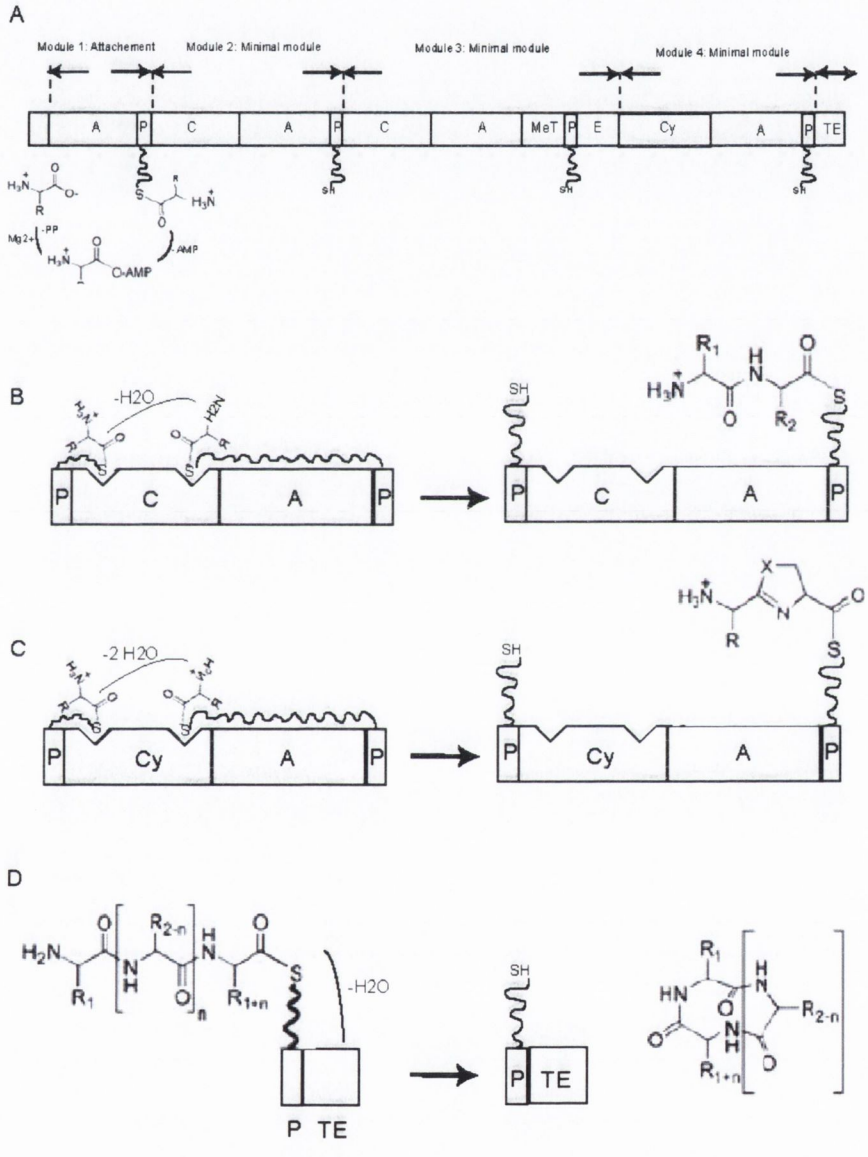


Figure 1.2 Modular organization of NRPS (Non-Ribosomal Peptide Synthetases). A. Module 1 is responsible of the upload of the first amino acid. Module 2 is the minimal structure and contains (A): adenylation domain that activates the amino acid, (PCP) is the peptidyl carrier domain responsible of the attachment of the amino acid as a thioester to 4'-phosphopantetheine; and (C) is the condensation domain involved in the formation of the peptides bonds with the precedent amino acid. Module 3 contains structures involved in modifying the added amino acid, the Epimerization (E) and Methylation (MeT) domains. Module 4 possess a cyclization domain (Cy) instead of the condensation domain (C) that is responsible of the cyclization of the amino acid and its condensation. The ThioEsterase (TE) domain adjacent to the last module permits the release of the peptide chain. Domain (A) catalyzes the activation of an amino acid by hydrolysis of an ATP molecule to produce a corresponding adenylate. Following this domain (A) attaches the intermediate onto the activated (PCP) domain. B. Domain (C) catalyses the bond between two amino acids. C. Domain (Cy) permits the cyclization of the amino acid and its condensation onto the peptide chain with a peptide bond. D. The (TE) domain is responsible of the release of the peptide by hydrolysis of the thioeseter.

1.3.2.3 Hybrid PKS-NRPS enzymes

Hybrid genes were only recently identified in fungi (Kroken et al. 2003; Bohnert et al. 2004; Song et al. 2004). The identified hybrids are modular systems, with the PKS and NRPS modules being found in all possible configurations: NRPS followed by PKS, the opposite, or alternation between the respective modules of PKS and NRPS (Figure 1.3). The modules of a hybrid are not different to those found in PKS and NRPS independently. The KS domains are the only exception. The KS domain sequence is less conserved between PKS-NRPS hybrids and the KS domain of the PKS and NRPS than are the other domains (Du et al. 2001; Du and Shen 2001; Du et al. 2003). The only hybrid genes that have been cloned to date are *FusS*, implicated in fusarin production (Song et al. 2004), *Ace1* for an unidentified metabolite (Bohnert et al. 2004), *eqiS* for equisitin (Sims et al. 2005), and LNKS, implicated in lovastatin production (Kennedy et al. 1999). All the above enzymes share the property of being type I iterative PKSs that possess only one NRPS module at their C-terminus (Figure 1.3). The identified metabolites possess a long carbon chain and one amino acid. It is proposed that the NRPS catalyses the link between the polyketide and the amino acid (Song et al. 2004). It remains necessary to clone other hybrids to better understand how they function.

1.3.2.4 Terpenoids

Two very well studied terpenoids are trichothecenes and gibberellins, both produced by *G. fujikuroi*. Terpenes are made from isoprene units, that are derived from dimethylallyl diphosphate and isopentenyl diphosphate condensation, produced by the mevalonate pathway (Townsend 1997; Keller et al. 2005). An alternative pathway involving methylerythritol phosphate exists in bacteria and plants, but has not been yet described in fungi (Rohmer et al. 2004; Dudareva et al. 2005). This condensation gives rise to geranyl diphosphate, a composite of the monoterpenes. The addition of another isopentenyl diphosphate gives rise to

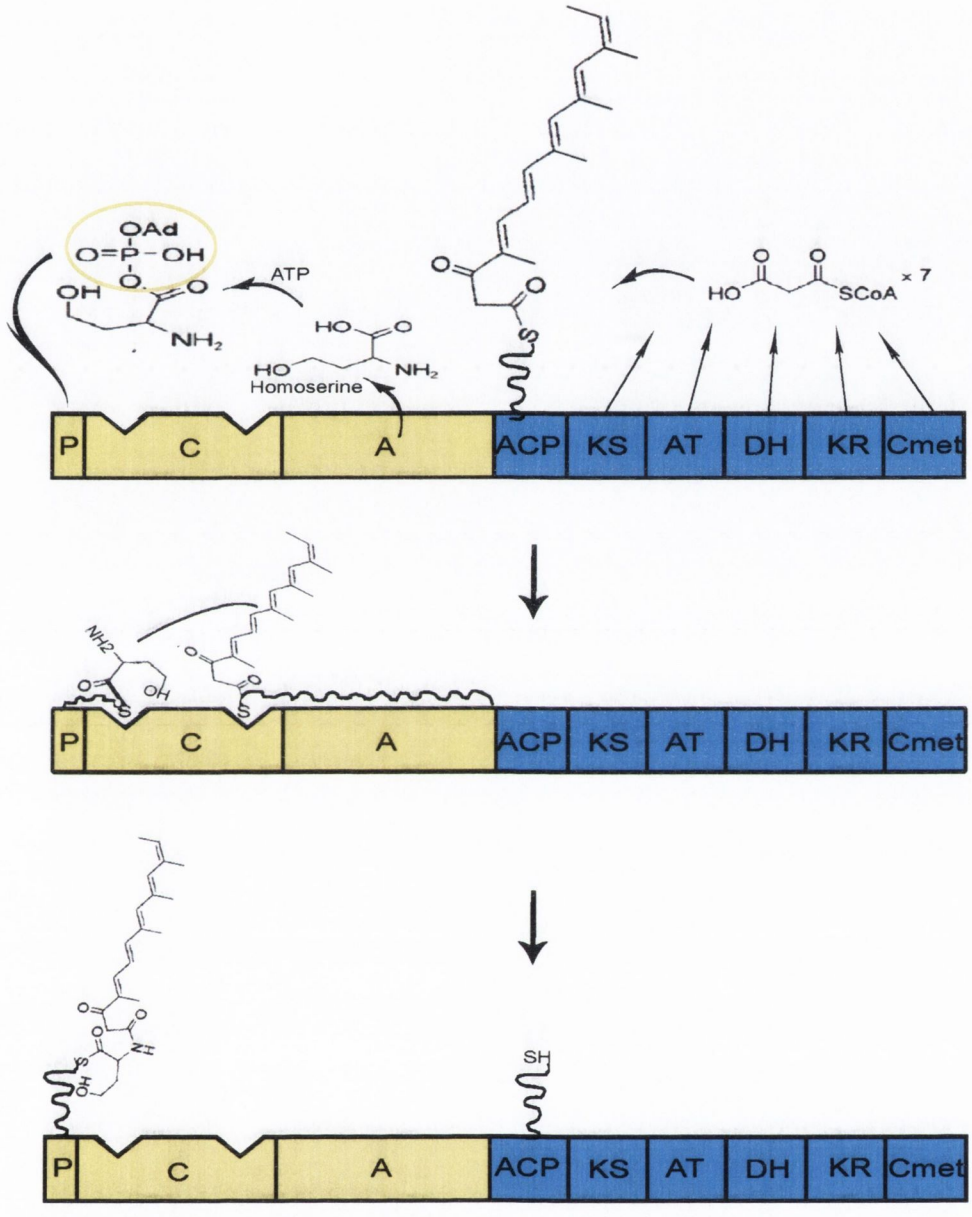


Figure 1.3 Example of a proposed biosynthesis pathway of a hybrid PKS-NRPS gene producing the Fusarin C produced by *Fusarium moniliform* and the fixation of the homoserine on the polyketide chain. The polyketide is synthetized by the PKS module of the FusS enzyme. The (PCP) domain of the NRPS domain is uploaded with the homoserine activated by the (A) domain, the (C) domain catalyses the condensation of peptide bond between the two molecules. See (Song et al. 2004) for more detail.

farnesyl diphosphate, a major component of steroids and sesquiterpenes, like trichothecenes. The geranylgeranyl diphosphate derived from a new condensation of isopentenyl diphosphate is used for synthesis of carotenoids and diterpenes.

These different precursors are at the root of the structural diversity of terpenes. The following steps will determine the structure of the metabolite. Particularly important are the terpenes cyclases, which are key enzymes in the biosynthesis pathway. For example, the first step in trichothecene biosynthesis is the cyclization of the farnesyl diphosphate to give rise to trichodiene. This reaction is catalyzed by Tri5, a trichodiene synthase (Hohn and Beremand 1989; Hohn and Desjardins 1992).

1.3.2.5 Alkaloids

Dimethylallyl diphosphate molecules are precursors for the synthesis of alkaloids. The biosynthetic pathway of one group of alkaloids – the ergot alkaloids – was determined in *Claviceps*. The first step is catalyzed by prenyl transferase, which binds the dimethylallyl to the amino acid tryptophan (Tudzynski et al. 1999). Other amino acids such as tyrosine can also be precursors. The dimethylallyl tryptophan then goes through a N-methylation followed by a series of oxidations that give rise to a range of intermediates, LSD (lysergic acid) being one of these intermediates. The addition of a non ribosomal tripeptide and other successive modifications give rise to a variety of ergopeptides (Haarmann et al. 2006).

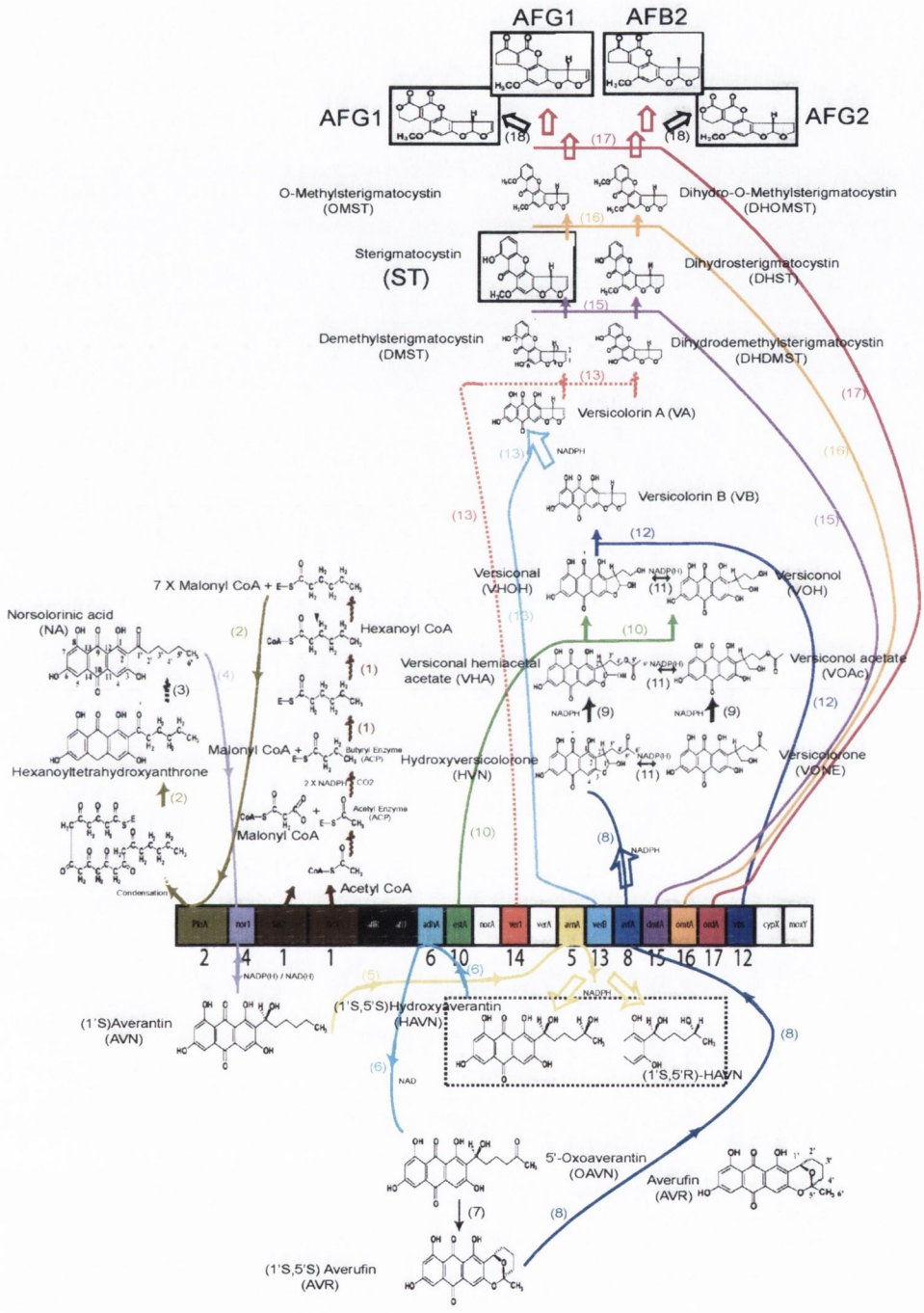


Figure 1.4 This figure is inspired from (Yabe and Nakajima 2004) and represents 18 of the reactiona in the aflatoxin and sterigmatocystin biosynthesis pathways starting from the acetyl CoA. Numbers indicate the sequential order of the reaction on the total 18 represented on this figure. The linked colored boxes are the aflatoxin clustered genes. Colors indicate a given reaction linked to a given enzyme in the cluster; black genes are regulatory genes AFLJ and AFLR; white genes are those that have not yet been linked to a reaction. The link can be an unconfirmed reaction shown in the figure with dashed arrows; open arrows a reaction catalyzed by a microsome enzyme; solid arrows a reaction catalyzed by a cytosolic enzyme. Black numbers and arrows indicate enzymes that are yet unidentified.

1.3.3 Genes in secondary metabolite pathways

At the heart of an SM synthesis pathway – and what will determine its class – are the PKS, NRPS, prenyl transferase or terpene cyclase enzymes. These enzymes synthesize the backbone of the metabolite, but the molecule undergoes extensive additional chemical reactions such as condensation, methylation and oxygenation, before the final SM product is completed. These additional steps are carried out by enzymes such as cytochrome P450 oxidoreductases, methyltransferases, and enoyl-reductases. Transporters of the MFS (major facilitator superfamily) or ABC (ATP-binding cassette) families are also often involved in SM pathways. These transporters can function either to secrete the final SM out of the cell (*e.g.*, *Tri12*, implicated in the efflux of trichothecene), or to protect the fungus against the newly synthesized metabolite, such as *Fum19* in *Fusarium verticillioides*, *sirA* in *Leptosphaeria maculans*, and *IvrA* in *Aspergillus nidulans*, which contribute to self-protection against fumonisin, sirodesmin, and lovastatin respectively (Hutchinson et al. 2000; Proctor et al. 2003; Gardiner et al. 2005a). Figure 1.4 illustrates the complexity of the sequential chemical reactions coded by the genes in the aflatoxin and sterigmatocystin biosynthetic pathways. The whole pathway is often regulated by a transcription factor of the $Zn(II)_2Cys_6$ type zinc-finger. which is specific to fungi.

This complex and yet neat process, in which each gene is responsible for a given step in the production of the SM, is striking (Figure 1.4). But much more striking is that the genes responsible for the pathway of biosynthesis and export of a particular SM, as well as transcriptional regulation of that pathway, are very often found in well-defined, contiguous regions of the genome called SM gene clusters.

1.4 Physical clustering of genes on chromosomes

1.4.1 Background

Physical clustering of certain types of genes has been observed at various degrees of linkage in practically all species, prokaryotic and eukaryotic. In prokaryotes, genes in the same metabolic pathway are often tightly linked into operons. Such a tight linkage of genes that function in the same pathway is much rarer in eukaryotes, especially if tandemly duplicated genes are disregarded. Only a few examples are known in animals, from nematodes and vertebrates (Graham et al. 1989; Zorio et al. 1994; Blumenthal 1998; Anonymous 1999; Blumenthal et al. 2002; Blumenthal and Gleason 2003; Sambrook et al. 2003; Field and Osbourn 2008). An increasing number of studies have revealed that although genes in eukaryotes do not show the same degree of clustering observed in prokaryotes, nonetheless gene order is not random with respect to function or expression.

In *S. cerevisiae*, for example, it has been shown that consecutive gene pairs have a higher level of co-expression than other genes (Cohen et al. 2000; Kruglyak and Tang 2000). More importantly, it has been shown that essential genes are highly clustered, independently of the clustering of co-expressed genes and of tandem duplications (Pal and Hurst 2003; Fischer et al. 2006). Clustering of genes in the same metabolic pathway has also been reported in *S. cerevisiae* (Lee and Sonnhammer 2003; Wong and Wolfe 2005; Hall and Dietrich 2007). The genomes of filamentous fungi appear to show a greater degree of clustering of genes with related functions, although this difference has not been quantified. As well as the clustering of genes involved in SM biosynthesis (Keller and Hohn 1997) which is the main focus of this thesis, filamentous fungi also show physical clustering of genes involved in catabolic pathways for the utilization of nutrients, such as quinic acid utilization (Patel et al. 1981; Schweizer et al. 1981), proline utilization (Hull et al. 1989), ethanol utilization (Felenbok 1991; Fillinger and Felenbok 1996), and nitrate assimilation (Johnstone et al. 1990).

Domains of similarly expressed genes have also been shown in animal genomes: in *Drosophila* (Spellman and Rubin 2002) with large domains of similarly expressed genes; in *C. elegans* with the co-expression of adjacent gene pairs (Lercher et al. 2003); in human with the clustering of highly expressed genes (Caron et al. 2001; Versteeg et al. 2003); and more generally in eukaryotes with the clustering of genes that have similar and/or coordinated expression (Hurst et al. 2004). Further, it has been shown that some categories of genes tend to cluster more than by chance: muscle specific genes tend to cluster in *C. elegans* (Roy et al. 2002); immunogenic and housekeeping genes are not randomly distributed in the mouse genome (Williams and Hurst 2002); and the human genome also shows clustering of tissue-specific, immunogenic and housekeeping genes (Bortoluzzi et al. 1998; Lercher et al. 2002; Megy et al. 2003). The clustering of pathway members is also reported in *C. elegans*, *D. melanogaster*, and human (Lee and Sonnhammer 2003). Nematodes take clustering a step further in that the clustering resembles a prokaryote operon, in that many clustered genes are transcribed polycistronically (Zorio et al. 1994; Blumenthal 1998; Blumenthal et al. 2002; Blumenthal and Gleason 2003). Finally, similar results are also found in plants, with the clustering of co-expressed genes (Williams and Bowles 2004), and genes involved in SM biosynthesis (Qi et al. 2004; Field and Osbourn 2008)

Further, it has been shown that genes from the same Gene Ontology classification (Fukuoka et al. 2004) or from the same metabolic pathway (Lee and Sonnhammer 2003; Wong and Wolfe 2005; Hall and Dietrich 2007) tend to cluster to a greater extent than expected by chance. Despite the results described above, relatively little is known about the clustering of functionally related genes in many taxa, since the definition of "functionally related" remains ambiguous. However, it seems beyond doubt that the filamentous fungi in particular show evidence of non-random gene order in many parts of their genomes. A striking feature of these genomes, the extent of which was only realized in the past few years, is the clustering of many genes implicated in SM biosynthesis, such as: aflatoxin (Trail et al. 1995; Yu et al. 1995), sterigmatocystin (Brown et al. 1996b), trichothecene (Brown et al. 2004), penicillin (Diez et al. 1990; Montenegro et al. 1990), and fumonisin (Proctor et al. 2003); and in catabolic pathways some of which were mentioned at the start of this section (Keller and Hohn 1997).

Relatively little is known about the extent of physical clustering of functionally related genes in filamentous fungi because, so far, no genome-wide statistical analyses have been carried out. There is, however, considerable anecdotal evidence that gene order is distinctly non-random in filamentous fungi for some sets of genes. The many large clusters of SM biosynthesis genes are the most striking examples, but in addition there have been many anecdotal reports of the clustering of other types of genes. Taking only the model organism *Aspergillus nidulans* as an example, there are reports of the physical clustering of (i) developmental genes (Orr and Timberlake 1982; Boylan et al. 1987); (ii) the clustering of co-expressed genes (Gwynne et al. 1984); and (iii) genes in pathways for the utilization of low molecular weight nutrients, for instance nitrate assimilation (Johnstone et al. 1990; Chang et al. 1996; Amaar and Moore 1998) and proline utilization (Hull et al. 1989). Although we cannot estimate what proportion of genes are located within such clusters, our impression is that the proportion is significantly higher for filamentous fungi than for the *Saccharomyces* group of hemiascomycetes.

This thesis will concentrate on a particular class of gene clusters that are involved in the synthesis of SMs, in several species of filamentous fungi.

1.4.2 Clustering of secondary metabolite clusters in filamentous fungi

Clustering of metabolic pathway genes in prokaryotes is a well-recognized characteristic of this kingdom. The first reports of what are now known as operons in prokaryotes date back to the late 1950s. Demerec and co-workers showed that genes involved in the biosynthesis of the same amino acid are often closely linked in *Salmonella* and *Escherichia coli* (Demerec and Ozeki 1959). Other example of operons followed this report, including operons for synthesis of histidine (Ames et al. 1963), leucine (Margolin 1963), and tryptophan (Somerville and Yanofsky 1964). Searches for similar instances of clustering in eukaryotes were made in the ascomycete fungus *Neurospora crassa*. In contrast to bacteria, many of the genes in the same pathways were found to be scattered in *N. crassa*, including the

leucine (Burns et al. 1963; Gross et al. 1963; Jungwirth et al. 1963) and tryptophan (Lester 1963) pathways. However, the histidine pathway was found to be partially clustered: out of the total of nine genes involved in histidine synthesis, three are clustered at the *N. crassa hist-3* locus (Ahmed et al. 1964; Horowitz 1965).

It was only in 1981 that further examples of the linkage of genes into clusters in filamentous fungi were found by Schweizer and co-workers (Patel et al. 1981; Schweizer et al. 1981), who cloned the genes involved in quinic acid catabolism in *Neurospora crassa* and found that they were in close physical proximity. This work was followed in 1989 by Arst, Scazzocchio and co-workers (Hull et al. 1989), who cloned the genes involved in L-proline catabolism in *A. nidulans* and found that these genes were also physically clustered.

The recognition that clustering is an important feature of fungal secondary metabolite pathways emerged over several years, with the molecular characterization of specific SM pathways. The clusters that led the way in this field of molecular characterization were the sterigmatocystin and aflatoxin clusters of *A. nidulans* and *A. flavus* respectively (Trail et al. 1995; Yu et al. 1995; Brown et al. 1996b). Despite the apparent improbability of having genes in the same pathway all occupying one locus, the repeated observation of the clustering of functionally related genes implicated in secondary metabolism eventually led to a clustered organization becoming recognized as the norm for SM gene organization in filamentous fungi (Keller and Hohn 1997).

Although genes for most SMs in the four main classes (see Section 1.3) are organized into clusters, there are however some exceptions. Examples include the genes coding for the melanin biosynthetic pathway, which are not clustered in *Magnaporthe grisea* or *Colletotrichum lagenarium* despite being clustered into a region of 30 kb in *Alternaria alternata* (Keller and Hohn 1997). Similarly the genes coding for T-toxin in *Cochliobolus heterostrophus* are not clustered (Kodama et al. 1999).

The sizes of SM gene clusters are variable and can range from 4 (or less, depending on one's definition of a cluster) to 25 genes, with the largest being the sterigmatocystin-aflatoxin cluster with 25 or 24 genes (Yu et al. 1995; Brown et al. 1996b). Other large clusters are those for lovastatin and sirodesmin with 18 genes (Kennedy et al. 1999; Gardiner et al. 2004), ergot alkaloid with 17 genes (Haarmann et al. 2006), fumonisin with 15 genes (Proctor et al. 2003), trichothecene with 12 genes (Kimura et al. 2003; Brown et al. 2004), and aurofusarin with 11 genes (Malz et al. 2005).

1.4.3 Why do fungi synthesize SMs?

To try to understand the benefit of having genes in secondary metabolism physically clustered, it is important to first understand the advantage that these metabolites bring to the species that make them, because the persistence of SM pathways depends on the selective advantage that they confer to the host organism. Little is known about reasons why fungi synthesize SMs, but they can be presumed to confer some selective advantage, in at least some environmental conditions. Our relative lack of knowledge about the natural functions of fungal SMs contrasts with the situation for bacteria and plants, for which some SMs have been shown to contribute to defense and to inter-organism communication (Osbourn 2003). It is thought that the advantage of SMs to fungi is of an ecological nature, to increase the virulence of the fungus towards plants, or to be used as weapons against other competing organisms. A few proven instances support these ideas, as described below.

In very few cases, an SM has been shown to correlate with virulence. For example, the SM trichothecene deoxynivalenol (DON, made by several fungi of the genus *Fusarium*) has been shown to be important in virulence and propagation in the host plant tissues (Proctor et al. 1995; Bai et al. 2002). DON suppresses plant defenses (Jansen et al. 2005). Another toxin that has been implicated in virulence is melanin (Henson et al. 1999). Melanins are dark brown pigments produced by many fungi during spore formation for deposition in the cell wall. Their best-known role is in oxidative stress resistance. Melanin-defective strains

are less viable, which in consequence affects their development and virulence (Henson et al. 1999). For example in *Colletotrichum lagenarium*, melanin is necessary for penetration into plants (Perpetua et al. 1996). In this species melanin has been associated with appressorium formation. Appressoria are infection structures required for host penetration, impairment in their formation reduces virulence. In *Alternaria alternata*, melanin deposition is involved in spore development (Kawamura et al. 1999).

Another example showing the virulence property of some toxins are the Host-Specific-Toxins (HST) (Otani 2000). The capacity to produce these phytotoxins determines the capacity of the species to attack a certain host, for example, *Cochliobolus victoriae* produces victorin that can only infect oat, while *C. carbonum* produces HC-toxin that can only infect corn (**Scheffer** et al. 1967). Another example includes the pyrichalasin H produced by *M. grisea* and that is specific to the millet species *Digitaria* (Tsurushima et al. 2005).

For fungi that do not colonize plants or other hosts, an alternative explanation for the existence of SMs could be that they are important in the colonization of a new ecosystem and thus the capacity to compete with other micro-organisms already occupying this niche. It is well known, for example, that penicillin is an anti-bacterial agent. More direct evidence comes from the discovery that DON, besides inhibiting the plants' defense (Proctor et al. 1995; Bai et al. 2002; Jansen et al. 2005), can also inhibit the *nag1* chitinase gene of *Trichoderma atroviride* (Lutz et al. 2003). Chitinase is an enzyme that degrades chitin, an important component of fungal cell walls (Duo-Chuan 2006). The idea is that DON is synthesized by several fungi of the genus *Fusarium* to protect its own cell wall from attack by chitinases secreted by other fungi such as *T. atroviride* (Lutz et al. 2003).

1.4.4 Why do SM genes cluster? Is there a unifying explanation for the clustering of SM biosynthesis genes in filamentous fungi?

In Section 1.4.3 above we showed that ability to produce SM compounds has selective advantages. But synthesis of the SM could occur even if the genes were scattered around the genome, so why are the genes clustered? Many tentative hypotheses to explain why these clusters exist and are conserved in such a way have been circulating in the fungal community (Walton 2000; Keller et al. 2005; Zhang et al. 2005). Many of the ideas are similar to those that have been discussed with regard to the origins of gene clusters in bacteria and other organisms. The main difference between the bacterial and fungal systems is that the bacterial gene clusters are usually (though not always) organized into operons, with only one promoter per cluster. In the fungal SM clusters, each gene has its own promoter.

Ideas about the evolution of SM gene clusters can be classified into two groups, which I have named the inheritance model and the selectionist models, as outlined below. In discussing these ideas, it is important to distinguish between two different evolutionary processes: the initial *formation* of a gene cluster, and the subsequent *maintenance* of the genes in this clustered organization. There is debate about the relative roles of natural selection versus random chance, in both of these processes.

1.4.4.1 The inheritance model

The inheritance model postulates that the observation that many SM gene clusters exist in fungi is simply due to inheritance. In other words, all the current clusters are descended (by vertical inheritance and cluster duplication) from one ancient SM gene cluster that existed in a fungal ancestor. This model does not address the question of how the original cluster was formed (it is agnostic as to whether selection was involved), but it asserts that the clustered organization of SM biosynthesis genes seen in modern species is simply a reflection of their history. It asserts that the clustered organization of the genes in present-day fungal genomes is not being maintained by natural selection, and hence that a clustered

organization does not currently have any selective advantage over a non-clustered (scattered) organization. This point of view was proposed by Horowitz (Horowitzn 1945; Horowitz 1965); reviewed by Lawrence and Roth (1996).

The first part of the inheritance model – the idea that many SM gene clusters have been formed by duplication of other clusters – is supported by extensive observations. All SM gene clusters have many genes in common. Most clusters are composed of a backbone gene such as a PKS or NRPS, modifying genes such as dehydrogenases, and transporter genes (see Section 1.3.3), giving the impression of a common origin. There are also many observations of relatively recently duplication of SM gene clusters. For example, the aflatoxin cluster is partly duplicated in *A. parasiticus* (Cary et al. 2002; Chang and Yu 2002), and the Ace1 cluster in *M. grisea* is the result of fusion of two clusters that were initially formed by tandem duplication (see Chapter 4).

The second part of the inheritance model – the proposal that the organization of SM genes into clusters is an accident of their history and is not being maintained by natural selection – is much more controversial. Many of the arguments that have been advanced in support of this idea are not in fact relevant to it: they are concerned with whether the SM *pathway* is being conserved by natural selection, not whether the *organization* of the genes into a cluster is being conserved.

Thus, the discovery that knockouts of many clusters do not seem to affect the fungus' virulence, growth, reproduction, or competitive ability, can be interpreted as support for the inheritance theory. If a cluster is important to the organism then we would expect that knocking it out would affect the fitness of the fungus, whereas if it is merely a vestige of an ancient cluster then it might not. Similarly, some fungal strains have been found to have a defective SM gene cluster and so cannot produce the SM, but this does not seem to have been selected against. An example is the AF36 strain of *A. flavus*, used as a bio-control agent to competitively exclude other aflatoxin-producing *Aspergillus* species (Ehrlich and Cotty 2004). The problem with these arguments is that the observations (both the laboratory knockouts and the discovery of polymorphisms for partly-deleted clusters) show that entire SM biosynthesis pathways can (sometimes) be

dispensable, whereas what we would really like to know is whether – in conditions where the SM biosynthesis pathway is advantageous or essential – there is any advantage to a clustered gene organization versus an organization where the genes are still present but are scattered around the genome. Finding that the whole pathway is unnecessary in some conditions (which may not be very relevant to real-life conditions in the wild) does not really tell us anything about whether natural selection cares whether the pathway's genes arranged into a cluster or scattered around the genome. By analogy, showing that the *lac* operon of *E. coli* is not essential in most growth conditions (it is only needed when cells are grown on lactose) does not tell us anything about whether there is an evolutionary disadvantage to having its genes (*lacZ*, *lacY* and *lacA*) scattered around the genome instead of arranged into an operon.

One way to test whether there is an evolutionary advantage to clustering *per se* is to investigate whether the clustered organization of SM biosynthesis genes has been conserved over long evolutionary timescales (between species) to a greater extent than we would expect by chance. For example, the gene clusters that currently code for aflatoxin in *A. flavus* and sterigmatocystin in *A. nidulans* (see Chapter 2) are descended from a common ancestor that existed at least 120 million years ago (Cary and Ehrlich 2006). Is this level of conservation greater than we would expect by chance? This question does not seem to have been addressed previously in the literature, so I tested it by carrying out a simple computer simulation for the aflatoxin/sterigmatocystin cluster. After identifying orthologous genes between *A. nidulans* and *A. flavus*, and identifying breakpoints of synteny, I calculate that 1345 out of 7068 orthologous intergenic spacers were the sites of evolutionary breakpoint since the divergence of these two species. Applying this number of breakpoints to a simulated genome leads to the result that the probability that a cluster the size of the aflatoxin cluster would still be intact after the genome sustained 1345 rearrangements is only about 1%. From this result I conclude that it is very unlikely that the aflatoxin/sterigmatocystin cluster would have escaped the machinery of rearrangement if no selection was maintaining the clustered organization of the genes.

This result shows that the inheritance model alone cannot explain the conservation of SM gene clusters: even if a cluster were to arise by duplication of an existing cluster, it would need some type of selection to maintain its clustered organization throughout evolution. Nevertheless, the idea that some SM gene clusters originated by duplication of other clusters certainly seems to be true. What is incorrect about the inheritance theory is the notion that history alone is sufficient to hold the clusters together; natural selection is required too. The duplication of clusters is likely to be similar, in evolutionary terms, to the duplication of individual genes (Lynch and Conery 2000): a duplicate cluster could be retained in the genome either by neofunctionalization (*e.g.*, a modification of the biosynthesis pathway so that a different SM molecule is made, although it should be noted that no empirical examples of the divergence of duplicate clusters are known) or subfunctionalization (*e.g.*, if a cluster that previously made two SMs undergoes duplication and each of the daughter clusters specializes into making one of the two previous products). If neither of these processes occurs, then it is likely that the cluster will become pseudogenized.

1.4.4.2 Selectionist theory

The second theory, which I will refer to as "the selectionist theory", supports the idea that selection acts to maintain the clustered organization throughout evolution. The proposed sources of the selection are that clustering can be beneficial to the genome (co-regulation is an example of this case), or that the clustering is advantageous to the clustered genes (such as the selfish operon theory, giving the genes a higher chance of being functional in a recipient genome after horizontal transfer). Because this theory postulates that selection is maintaining the clusters, it is a simple step to invoke selection as a force in the origin of the clusters too.

There has been increasing evidence that natural selection plays a role in how genes are arranged along chromosomes in many eukaryotic genomes, including mammals (Singer et al. 2005) and yeast (Hurst et al. 2002). Co-expressed (Cohen et al. 2000), housekeeping (Lercher et al. 2002) and tissue specific genes (Roy et al. 2002; Megy et al. 2003) are all clustered to a greater extent than expected by

chance. The difficulty in proposing that natural selection can explain the clustering of SM biosynthesis genes in filamentous fungi has been due, for a very long time, to our lack of knowledge of what the benefit of clustering is, in the specific case of SM genes. Two possible benefits have been proposed and are discussed in the next section. But despite this lack of knowledge about its nature, the existence of a strong selective advantage to a clustered organization is indicated by the discovery of two examples of fungal gene clusters being created, over short evolutionary timescales, by the relocation of scattered genes into a single location in the genome. These two examples are the *DAL* gene cluster of *S. cerevisiae* (Wong and Wolfe 2005) and the fumonisin cluster of *Fusarium verticillioides* (Chapter 2).

In addition to the evidence that selection is involved in the formation of gene clusters, the same observations that argue against the inheritance model also argue in support of the selectionist theory. In particular, the conservation of the aflatoxin/sterigmatocystin cluster, described previously in Section 1.4.4.1, indicates a role of selection in the maintenance of SM biosynthesis clusters.

Several different proposals have been made concerning what type of selective advantage a clustered arrangement might have over a non-clustered one. In the sections below I review three of these ideas: (i) the selfish operon theory; (ii) the co-regulation theory; and (iii) the co-adapted gene cluster theory.

1.4.4.2.1 Selfish operon theory

Under the selfish operon theory, clustering benefits the operon. This hypothesis was first proposed in the bacterial literature (Lawrence and Roth 1996), but it is equally applicable to SM gene clusters in fungi. According to this hypothesis, clustering benefits the genes in the operon because their tight linkage increases the chances of successful horizontal transfer to another genome. Having all the genes in a pathway at the same locus dramatically increases the probability that a horizontal transfer event could successfully transfer a complete metabolic pathway from one genome to another. In a bacterial operon, co-transcription of the genes from a single promoter may also increase the chances of successful

activation of the whole pathway after horizontal transfer. Although fungal SM gene clusters do not constitute operons, they share the same characteristics of co-regulation and co-expression, and they very often contain all the steps implicated in synthesis of one metabolite. The strongest supporting argument for the "selfish" argument in fungi is the patchy phylogenetic distribution of SM gene clusters. We often observe that the same cluster is present in some very distantly related species, while closer species do not possess it. Such a distribution points to horizontal transfer between species. Presence of an ancestral cluster followed by losses in multiple lineages could also provide an explanation, but the patchier a cluster's phylogenetic distribution is, the less parsimonious this argument becomes.

Under this theory, horizontal gene transfer (HGT) is beneficial for functionally-coupled but weakly-selected genes. One will expect to see that the majority of coupled genes are non-essential. It was on this point that the "selfish" theory was rejected as a sole explanation for the conservation of such clusters in bacteria. Hurst and colleagues showed that in *E. coli*, genes with essential functions are more abundant in operons, concluding that the selfish operon theory cannot provide a general and unifying model to explain clustering (Pal and Hurst 2004).

There has been little previous evidence for HGT in the evolution of fungal SM clusters. In Chapter 3 I report an example of HGT (the Ace1 cluster) between Sordariomycete and Eurotiomycete fungi. A second example, of the ETP cluster in filamentous ascomycetes, was reported by Patron and co-workers (Patron et al. 2007). The lack of widespread evidence of HGT of many SM gene clusters among many fungi suggests, however, that it could be the sole explanation for why SM genes are maintained in clusters.

1.4.4.2.2 Co-regulation theory

This theory proposes that the genes in the clusters remain linked because the clustering is advantageous to the individual, because it allows the genes to be co-regulated (Keller and Hohn 1997). Indeed, the ability to coordinate the expression of every gene in an SM pathway, and to have quick induction or repression of

gene expression, is likely to be important, particularly for metabolites such as toxins. But the strongest argument for this theory, and against the selfish operon theory, is the constant recruitment of new genes into these pathways and the loss of other genes. If such cluster was only retained for the cluster's benefit, we will not expect to see selection on the cluster to change by recruiting new genes. A great example of this process is the aflatoxin and sterigmatocystin clusters, which share 20 genes in common but differ by at least five in each (Cary and Ehrlich 2006). Many other examples of such differences exist, for example the sirodesmin cluster that is present in many very distantly related genomes but differs dramatically in the number of shared and non shared genes (Gardiner et al. 2004). This argues favorably for selection on clusters to adapt to the needs of the individual, and thus argues for selection towards the benefit of the species as opposed to the genes in the cluster.

The hypothesis that the evolutionary advantage of a clustered organization for SM biosynthesis genes received strong support in 2004 from Bok and Keller's discovery of the global regulatory gene *laeA* in *A. nidulans* (Bok and Keller 2004). Orthologs of this gene have been since found in all Aspergilli, and also in more distantly related species, such as *M. grisea* and *N. crassa* (Keller et al. 2006). Deletion of *laeA* in *A. nidulans* blocks the expression of many SM gene clusters: sterigmatocystin, lovastatin, pigments, gliotoxin and penicillin (Bok and Keller 2004; Keller et al. 2006). Similarly, microarray experiments on gene expression in a *laeA* deletion strain of *A. fumigatus* revealed that 13 of its 22 SM gene clusters are regulated by LaeA. The existence of such a universal regulator of SM gene expression might provide a good explanation for the arrangement of these genes into clusters. Although the functions of LaeA are not yet fully characterized, it is speculated to be involved in chromatin modification such as defining the boundaries of euchromatic and heterochromatic DNA for the cluster (Keller et al. 2005; Keller et al. 2006), and thus might be involved in both gene repression and activation in heterochromatic and euchromatic regions of eukaryotic chromosomes (Lee et al. 2005). Other genes artificially inserted in the cluster are not expressed in LaeA mutants, which suggests a role in chromatin modification (Bok et al. 2006). Therefore, having the genes in one locus facilitates a general regulation by LaeA. This provides a higher level of regulation, above

the functional of local regulator genes such as *aflR* (in the aflatoxin cluster) which are themselves located within the SM gene clusters.

A role of chromatin in regulating SM gene clusters is also suggested by the fact that the clusters are often found in condensed islands, distant from all other neighboring genes. This common feature might reflect the importance of expressing the genes only in very specific moments of the fungus' lifecycle; spillover of the cluster's expression to other neighboring genes, or vice versa, might have adverse effects. The production of certain metabolites is tightly linked to morphological development of the fungus. For example, some nonaflatoxigenic variants of *A. parasiticus* are incapable of sporulating, and the authors suggested that regulation of aflatoxin synthesis and conidiogenesis may be interlinked (Kale et al. 1996). Most secondary metabolites are produced after the fungus has completed its initial growth phase and is beginning a stage of development represented by the formation of spores (Calvo et al. 2002). For example mycotoxins in the *Aspergillus* species are produced after sporulation (Reiss 1982), as is zearalenone in *Fusarium graminearum* (Wolf and Mirocha 1973), and linoleic acid in *A. nidulans* induces sporulation (Champe et al. 1987; Champe and el-Zayat 1989; Calvo et al. 2001). In plants it has been shown that many SMs are highly tissue specific and under strict developmental control. For example avenacins are localized in the epidermal cells of the root tip (Osbourn et al. 1994), and AsbAS1 is expressed specifically in this cell layer (Haralampidis et al. 2001). Clustering has the potential to facilitate coordinate regulation of expression of genes at the chromatin level for a specific time period in which the pathway is expressed.

Strict control and containment of SMs and their pathway intermediates during synthesis is likely to be important and thus may explain part of the reason for the clustering. Disruption of the gene cluster may lead to failure to produce protective chemicals and could also result in the accumulation of deleterious intermediates (Gierl and Frey 2001).

1.4.4.3 Coadapted gene complexes

A third hypothesis about the origins of clusters of co-functional genes is the hypothesis that epistatic section (Nei 1967) can form a coadapted gene complex. According to this hypothesis, if two loci are polymorphic (say we have alleles *A1* and *A2* at locus *A*, and alleles *B1* and *B2* at locus *B*), and if one allele at each locus works better with a particular allele at the other locus (*e.g.*, the genotypes *A1 B1* and *A2 B2* are fitter than *A1 B2* or *A2 B1*), then there will be selection in favour of relocating genes *A* and *B* close to each other (Nei 1967; Nei 2003; Hurst et al. 2004). If they become so close that there is strong linkage disequilibrium between them, then each pair will effectively act as a single allele (*A1-B1* and *A2-B2*). This model has been used to explain the existence of gene clusters such as the human MHC, that are highly polymorphic.

1.4.4.4 Conclusion

In conclusion, the reason for the current lack of an answer to the important question "**Is there a unifying explanation for the clustering of SM genes in filamentous fungi?**" might be that several evolutionary mechanisms account for clustering of SM biosynthesis genes.

However, in this chapter I reject the inheritance theory, according to which clustering is not an adaptive property. In this thesis I show that selection acts on creating SM gene clusters; on gene rearrangements within a cluster; and that individual genes in clusters are under selection to adapt to niche changes, such as the differences observed in the genes in the fumonisin cluster (Chapter 2) and in *aflR* and *aflJ* from the aflatoxin cluster (data not shown). In Chapter 3, I show that the most common mechanism of cluster expansion is gene duplication, as demonstrated by the significantly low number of clusters in *N. crassa*, probably due to the repeat-induced point (RIP) mutations this genome is experiencing. I also show that new clusters can be formed by *de novo* assembly of previously scattered genes, such as the assembly of the fumonisin cluster (Chapter 2). I also reject the selfish operon theory as an explanation for the formation of SM gene clusters in filamentous fungi. However I support the idea that SM clusters (once

formed) can be transmitted by HGT, without it being the reason for the clustering as the selfish operon theory postulates (Chapter 4).

I conclude this chapter by putting forward the idea that SM gene clusters act as a larger version of a gene: they can duplicate, diverge, pseudogenise and can thus be lost differentially in different lineages. But SM gene clusters also have a much greater freedom to rearrange, grow, and shrink, than genes do.

Chapter 2

Evolutionary origins of the fumonisin secondary metabolite gene cluster in *Fusarium verticillioides*

This chapter is based on a manuscript that is currently in preparation (authors **N. Khaldi** and K.H. Wolfe).

2.1 Abstract

The secondary metabolite gene clusters of euascomycete fungi are among the largest known clusters of functionally related genes in eukaryotes. Most of these clusters are species-specific or genus-specific, and little is known about how they are formed during evolution. We used a comparative genomics approach to study the evolutionary origins of a secondary metabolite cluster that synthesizes a polyketide derivative, namely the fumonisin (FUM) cluster of *Fusarium verticillioides*. We identified orthologs in other euascomycetes of 13 of the 15 genes in the FUM cluster. These orthologs are not clustered in the other species and many of them are located in conserved genomic contexts. The *FUM* cluster became assembled by moving these genes to a new genomic site; orthologs of the genes flanking the 45kb FUM cluster in *F. verticillioides* are immediately beside one another in *F. graminearum*, *Magnaporthe grisea* and *A. nidulans*. This mode of cluster assembly is similar to that of the *DAL* cluster in *Saccharomyces*

cerevisiae. Moreover, we find that the FUM cluster genes are among the fastest-evolving genes in the *F. verticillioides* genome, and they have accelerated in this species since it diverged from *F. graminearum*.

2.2 Introduction

The order of genes along eukaryotic chromosomes is often assumed to be random, but there is growing evidence that the chromosomal position of some genes is maintained by natural selection (Hurst et al. 2002; Hurst et al. 2004; Singer et al. 2005), and that selection can sometimes operate to move genes to new locations (Qi et al. 2004; Wong and Wolfe 2005; Field and Osbourn 2008). Among the eukaryotes, many of the most notable examples of the physical clustering of genes with related functions occur in the filamentous fungi (Giles et al. 1985; Hull et al. 1989; Keller and Hohn 1997; Cary et al. 2001). The most striking fungal gene clusters are those involved in the synthesis of secondary metabolites such as aflatoxin (a 25-gene cluster in *Aspergillus flavus*), fumonisin (a 15-gene cluster in *Fusarium verticillioides*), and trichothecene (a 12-gene cluster in *Fusarium sporotrichioides* (Brown et al. 2004). Secondary metabolites are organic molecules that are not essential for the normal growth of the fungus but which function in host/pathogen interactions or other forms of communication or warfare between organisms. They are typically modified polyketides, terpenes, alkaloids or non-ribosomal peptides (Keller et al. 2005). Synthesis of these molecules requires many successive enzymatic steps, and the genes for these enzymes are almost invariably clustered together at a single genomic location. It has been suggested that physical clustering may allow the genes to be co-regulated by means of chromatin modification (Keller et al. 2005), or that it may facilitate horizontal transfer of intact clusters between species (Rosewich and Kistler 2000).

Since the discovery of secondary metabolite gene clusters, their mechanism of origin and assembly has remained a matter of speculation. The growing number of available euascomycete genome sequences now enables us to take a phylogenomic approach to the evolutionary origins of these clusters. In this study, we focus on two of the largest known secondary metabolite clusters, both of the

polyketide synthase type. We chose these two clusters simply because of their size. Fumonisins are mycotoxins produced by maize pathogens in the *Gibberella fujikuroi* complex, of which *F. verticillioides* and *F. proliferatum* are the most studied (*Gibberella* is the teleomorphic form of the genus *Fusarium*). The FUM genes are organized as a cluster of 15 genes (approximately 45 kb) in both species, though the location of the cluster differs between the two (Proctor et al. 2003; Waalwijk et al. 2004). The products of the FUM cluster genes include a polyketide synthase, fatty acyl-CoA synthases, and cytochrome P450 monoxygenases. Expression of the genes in the cluster, but not the neighboring genes, is induced under conditions when fumonisin is synthesized (Proctor et al. 2003). The ability to synthesize fumonisin has a patchy phylogenetic distribution across the genus *Fusarium*, due to the variable presence or absence of the FUM cluster among different isolates (Proctor et al. 2004).

We combined comparative genomics with phylogenetic analysis to investigate whether genes in the FUM cluster have orthologs in filamentous fungi that do not synthesize the mycotoxins, and so to study how these clusters were formed.

2.3 Methods

Our analysis was done using the completely sequenced genomes of the euascomycetes *A. nidulans*(Galagan et al. 2005), *F. graminearum* (MIT WI)(Cuomo et al. 2007), *M. grisea* (Dean et al. 2005), and *N. crassa* (Galagan et al. 2003).A set of 87,000 expressed sequence tags from *F. verticillioides* (Brown et al. 2005) was used for analysis of rates of sequence evolution in this species compared to *F. graminearum*. We were unable to use the sequence of the *F. proliferatum FUM* cluster (Waalwijk et al. 2004) because it is not public.

To identify orthologs of genes in the FUM and AFL/ST clusters, we first used each protein as a query in a BLASTP search against the NCBI nonredundant protein sequence database. Sequences giving hits with Expect (E) values of less than 1e-4 were retained and used for phylogenetic analysis. Each set of proteins was aligned using ClustalW (Thompson et al. 1994) and poorly aligned regions

were removed using Gblocks (Castresana 2000). Maximum likelihood trees were constructed using PHYML (Guindon and Gascuel 2003) with the JTT amino acid substitution matrix and four categories of substitution rates. Bootstrapping was done using the default options in PHYML with 100 replicates per run. We used yn00 in the PAML package (Yang 1997b) to determine the values of the non-synonymous substitutions per site (*dN*). A monophyletic clade with a bootstrap of 70 or more was required as a definition of orthologous genes represented in Figure 2.1.

The probability that multiple independent disassembly events could explain the genomic locations of FUM cluster orthologs in other species (Figure 2.1) was estimated as follows. We assume that the number of possible intergenic regions into which a transposing gene could become integrated is 10^5 (the approximate number of genes in each genome). If two independent transpositions of the same gene in two different lineages occur, the probability that their integration sites are the same is therefore 10^{-5}, assuming that every intergenic site is equally likely to become an integration site. For three independent transpositions, the probability that all three sites coincide is 10^{-10}. In Figure 2.1 there are three independent lineages: the *F. graminearum* branch, the *A. nidulans* branch, and the shared branch leading to *N. crassa* and *M. grisea*. For the orthologs of *FUM10*, *FUM16*, *FUM17* and *FUM18* we see conservation of ortholog locations on two of the three lineages, and for *FUM11* we see conservation on all three lineages. Hence, under the disassembly hypothesis, there were four independent cases of coincidence of two integration sites, and one case of coincidence of three integration sites, so the joint probability of all five cases is $(10^{-5})^4 \times (10^{-10}) = 10^{-30}$.

2.4 Results

2.4.1 Recent assembly of the FUM gene cluster

We used phylogenetic analysis to identity apparent orthologs of genes in or near the *F. verticillioides* FUM cluster genes in other filamentous fungal genomes. Orthology relationships between the genes in or near the *F. verticillioides* FUM

cluster and other euascomycete genes are summarized in Figure 2.1. We identified apparent orthologs (based on phylogenetic analysis) of 13 of the 15 FUM genes in

at least one other euascomycete species. These genes do not form clusters in the other genomes where they are present. A region of *F. graminearum* chromosome 1 (genes *FG00269–F00276*; Figure 2.1) contains orthologs of the five genes to the left of the FUM cluster (*F. verticillioides NPT1*, *WDR1*, *PNG1*, *ZNF1* and *ZBD1*) immediately adjacent to orthologs of the two genes to the right of the cluster (*F. verticillioides ORF21* and *MPU1*), with nothing in between them in *F. graminearum*. Similarly, in *M. grisea* the *ZBD1* ortholog is beside the *ORF21* ortholog, and in *A. nidulans* the *PNG1* ortholog is beside the *ORF21* ortholog. Thus, chromosomal sites orthologous to the FUM locus exist in *F. graminearum*, *M. grisea* and *A. nidulans*, but no orthologs of the FUM genes themselves are found at these sites (Figure 2.1). Instead, the orthologs of the FUM genes are scattered on different chromosomes of the genomes of these non-FUM producing species. In *F. graminearum*, for instance, the 10 orthologs of FUM genes dispersed across four chromosomes and none of them are located close to each other or to the FUM-flanking genes. Taking into account the phylogenetic relationship among these genomes (Figure 2.1), this arrangement indicates that either (*i*) the FUM cluster genes were inserted into a pre-existing genomic locus between *ZBD1* and *ORF21* in an event that occurred in a recent ancestor of *F. verticillioides*, after it had diverged from the other species; or (*ii*) the cluster existed in the ancestor of all five species represented in Figure 2.1, but was retained only in *F. verticillioides* and was lost in the other four genomes.

To assess whether the FUM cluster became assembled in *F. verticillioides*, or disassembled in the other genomes, we examined the genomic contexts around each of the FUM cluster orthologs in other species. For example, the *F. verticillioides* FUM cluster gene *FUM11* is orthologous to *FG07875* in *F. graminearum*, and to *MG03479* in *M. grisea* (Figure 2.1). We will refer to *FG07875* and *MG03479* as focal genes. When we examine the regions around these focal genes in the *F. graminearum* and *M. grisea* genomes, we find that some of the neighboring genes near them are also orthologs of each other. On one side *FG07874* (1 gene away from the focal gene) is an ortholog of *MG03474* (5

genes away), and on the other side *FG07864* (11 genes away from the focal gene) is an ortholog of *MG03480* (1 gene away from the focal gene). In Figure 2.1, only the focal genes are shown but these similarities of context are indicated by the

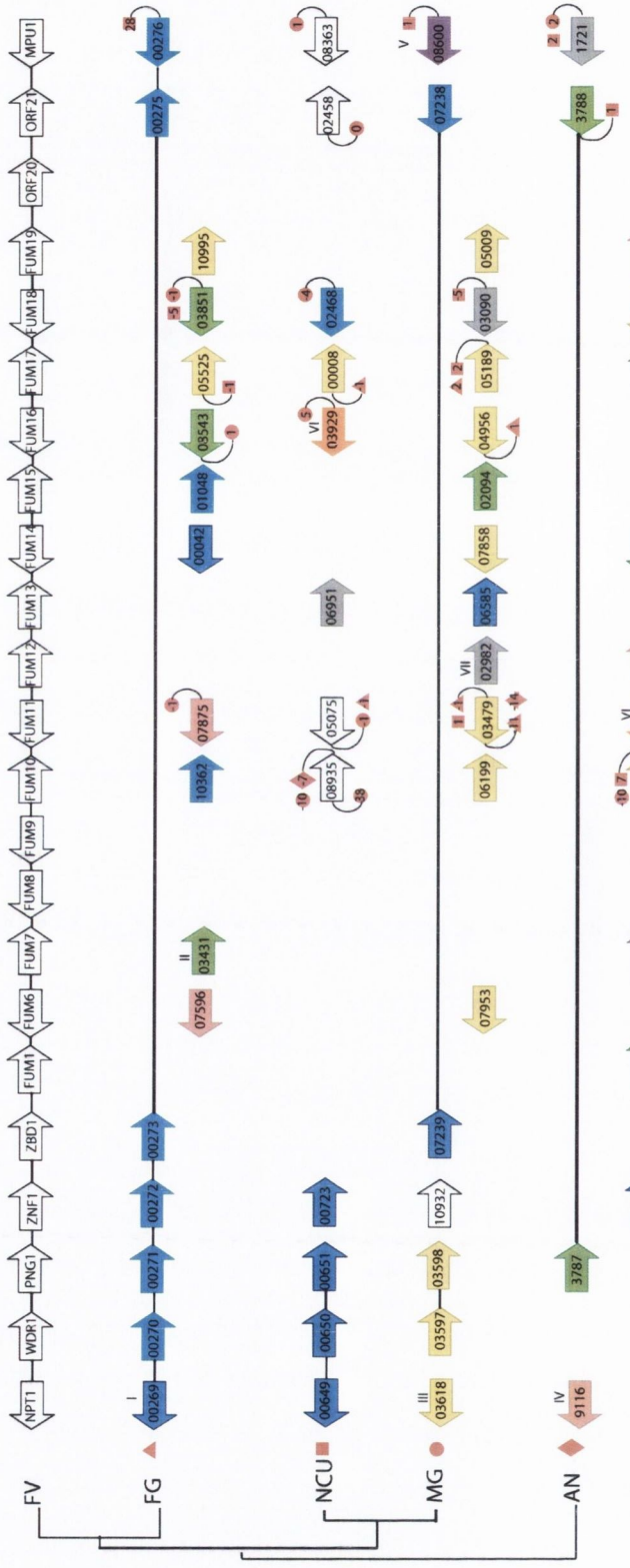


Figure 2.1 Representation of the fumonisin cluster and its flanking genes in *F. verticillioides* (FV). Columns are orthologs identified by phylogenetic analysis in *F. graminearum* (FG), *N. crassa* (NCU), *M. grisea* (MG), and *A. nidulans* (AN). Genes in the latter four species are identified by gene numbers from their genome projects. Different colors represent different chromosomes. The long lines in *F. graminearum*, *M. grisea* and *A nidulans* show that in those species, there is a site in the genome that corresponds to the FUM cluster location, but no FUM genes are present at that locus. Curved lines and numbers in orange symbols indicate conservation of the neighboring genes around FUM orthologs in *F. graminearum*, *N. crassa*, *M. grisea* and *A. nidulans*. For each gene show in the figure (the focal genes) we considered the two genes immediately next to it. If these genes have orthologs located <20 genes away from the orthologous gene in another species, a symbol indicates this fact. For example, the numbers –10 (in a circle) and –7 (in a diamond) connected to gene *NCU08935* indicate that the gene immediately after *NCU08935* (i.e. *NCU08936*) has an ortholog in *M. grisea* that is 10 genes away from *MG06199* (*i.e., MG06189*), and an ortholog in *A. nidulans* that is 7 genes away from *AN04397* (*i.e., AN04392*). Triangles, squares, circles and diamonds indicate relationships to *F. graminearum*, *N. crassa*, *M. grisea* and *A. nidulans*, respectively.

small numbers at the ends of the curved lines attached to the symbols for the focal genes *FG07875* and *MG03479*.

Overall, the orthologs of five FUM cluster genes (*FUM10*, *FUM11*, *FUM16*, *FUM17* and *FUM18*) show some degree of local synteny conservation among the four species that do not contain FUM cluster. In other words, each of these genes is in a conserved location among some of the four species, and these locations are not close to one another. Given the structure of the phylogenetic tree, these conserved synteny relationships make it far more parsimonious to argue that the FUM cluster became assembled in *F. verticillioides* than to argue that it was present ancestrally and became disassembled multiple times (a minimum of three independent disassemblies are required: one on the *F. graminearum* lineage, one in the common ancestor of *N. crassa* and *M. grisea*, and one on the *A. nidulans* lineage). The hypothesis of assembly requires that each gene transposed once, from an ancestral conserved location, to its current location in *F. verticillioides*. In contrast, the hypothesis of disassembly requires us to believe that that, for example, *FUM10* transposed out of the cluster three times, and that on each of these occasions the transposed gene landed at the same (orthologous) genomic site in each species, giving the spurious appearance that an ancestral location was conserved. According to the disassembly hypothesis, the similarity of genomic context around the orthologs of each FUM-cluster gene in other species is the result of multiple convergent transpositions to the same sites in each species. Since a transposing gene can presumably land almost anywhere in the genome, such a convergent scenario is extremely unlikely. We estimate that the probability that multiple independent transpositions of five genes could produce the result shown in Figure 2.1 is approximately $P = 10^{-30}$ (see Section 2.3 Methods for details) which makes the disassembly hypothesis untenable. This finding also rules out the possibility of HGT of the FUM cluster as an explanation of its origin in *F. verticillioides*. Since the FUM cluster is also found *F. oxysporum*, the closest relative of *F. verticillioides*, we conclude that the FUM cluster became assembled on the branch leading to the common ancestor of *F. verticillioides* and *F. oxysporum*, after its speciation from the *F. graminearum* lineage.

The FUM genes became relocated, from their conserved ancestral locations, to their current positions within the cluster. This mode of assembly is similar that inferred for the *DAL* gene cluster in *S. cerevisiae* (Wong and Wolfe 2005). The mechanism indicates that very strong natural selection was involved in formation of the cluster. The genes must have been sequentially relocated, with selection for each step. Relocation could involve either gene transposition (cut-and-paste mechanism), or gene duplication followed by loss of the original copy of the gene (copy-paste-delete mechanism).

2.4.2 Accelerated sequence evolution of the fumonisin cluster genes

If there has been strong selection for gene relocation, we might also expect to see accelerated rates of sequence evolution in *F. verticillioides FUM* genes. We tested this idea by comparing the extents of sequence divergence in genes of *F. graminearum* and *F. verticillioides* from their orthologs in an outgroup species, *M. grisea.* In order to be able to compare *FUM* genes to the rest of the genome, we assembled a set of 2303 contigs from *F. verticillioides* expressed sequence tags (ESTs) (Brown et al. 2005) that were reciprocal-best-BLAST hits to non-*FUM* genes in the *F. graminearum* genome sequence (Cuomo et al. 2007). After excluding genes with very high levels of nonsynonymous nucleotide divergence ($dN > 1$), which might indicate paralogy, we were left with 1939 non-FUM genes and 8 genes from the FUM cluster (*FUM8* and *FUM14* were omitted because they, too, have $dN > 1$).

The results show that the FUM genes are among the fastest-evolving genes in the genomes of both of the *Fusarium* species. For the comparison of *F. verticillioides* to *M. grisea*, the mean *dN* for *FUM* genes is about 10 times greater than that for other genes (Figure 2.2 B). In both of the *Fusarium* species, the *FUM* genes appear as outliers in the distribution of dN values (Figure 2.2 C,D). Moreover, the rate of *FUM* gene evolution is faster in *F. verticillioides* (where they are

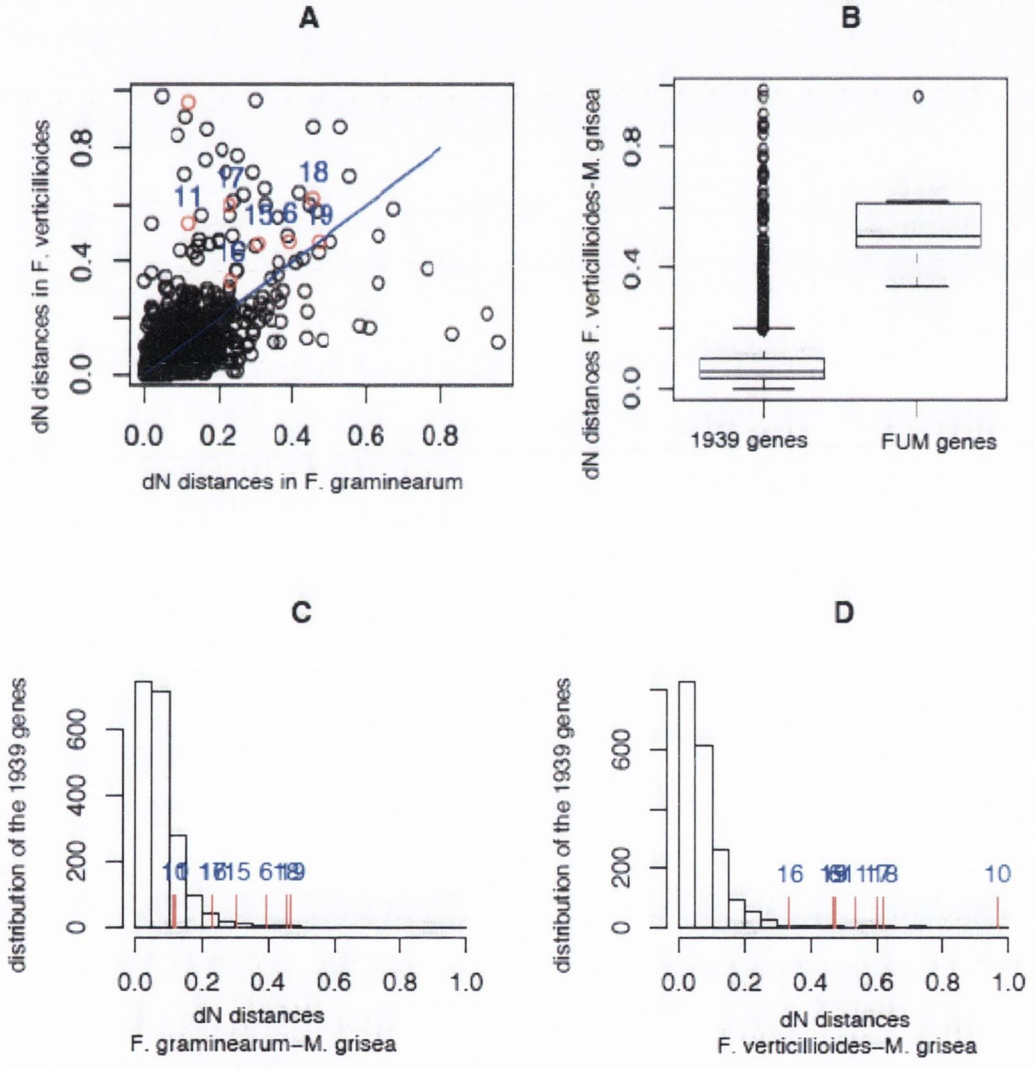


Figure 2.2 Sequence evolution of the fumonisin proteins. (A) Plot of the non-synonymous substitutions dN in *F. graminearum* versus *F. verticillioides*. Seqeunces evolving equally between both species should be on the blue line. Numbers indicate the points corresponding to FUM genes, e.g. "18" represents the Fum18 gene. (B) Boxplot of protein distances in the FUM genes versus 1939 other genes between *F. graminearum* and *F. verticillioides*. (C,D) Histograms of the distribution of dN distances between *F. graminearum* and *M. grisea* (C), and between *F. verticillioides* and *M. grisea* (D). The positions of the FUM genes are indicated by number.

clustered) than in *F. graminearum* (where they are not) (Figure 2.2 Figure 2.2 Figure 2.2 A). These results are suggestive of positive selection on *FUM* genes in

both of the *Fusarium* species, with further acceleration in *F. verticillioides* associated with the gene relocations that formed the cluster.

2.5 Discussion

In the literature three scenarios for the creation of a cluster have been described: horizontal of an existing cluster from one genome to another ((Patron et al. 2007; Khaldi et al. 2008) and Chapter 4); the duplication of an ancestral cluster (Chapter 4); and the *de novo* creation of a cluster from initially scattered genes that become relocated into one locus (Wong and Wolfe 2005). We find that the FUM cluster follows the third scenario, with a striking pattern of gene transposition to assemble the cluster. This result resembles our previous observations on the *DAL* gene cluster of *S. cerevisiae*, though it should be noted that the *DAL* genes code for a catabolic pathway (degradation of allantoin, a secondary nitrogen source) whereas the *FUM* genes are part of an anabolic pathway (secondary metabolite biosynthesis). A further difference is that we find evidence for a highly accelerated rate of sequence evolution in many of the genes in the FUM cluster (Figure 2.2), whereas the *DAL* genes did not show any acceleration. This result is suggestive of positive selection on the *FUM* genes, but we are not able to test directly for this using a dN/dS ratio test, because the synonymous divergence between *F. verticillioides* and *F. graminearum* is too large. Nevertheless, our results point to very strong selective forces operating on the FUM cluster during its recent evolution. Our lack of knowledge about what benefit the metabolite confers on the organism hampers our understanding of the selective purpose of this clustering. However, we can be almost certain that the reason behind the clustering is not simply to synthesize the metabolite, which is possible with scattered genes. It is more likely that the selective force involves selection for a tight co-regulation of gene expression, perhaps mediated by a LaeA-type universal regulator.

Competition, either between one fungal species and another, or between a fungus and a host species, is likely to result in strong selection on the secondary metabolite repertoire of filamentous fungal species. This arms race between living

species pressurizes the organism to create new chemical weapons, which are the products of new secondary metabolite gene clusters. It is relatively easy to envisage that neofunctionalization after gene duplication, or partial cluster duplication as appears to have happened in the origins of the Ace1 cluster (Chapter 4), could result in the production of a new secondary metabolite and so could be selectively advantageous. It is harder to understand why relocating genes, as has happened in the FUM cluster, can be evolutionarily advantageous. One possibility is that the mere act of relocating a gene can have the consequence of changing the end product of a pathway, because the expression of all the genes in a cluster is coordinated. For example, imagine that we have two secondary metabolite biosynthesis pathways, 1 and 2. If a cytochrome P450 oxidoreductase gene that originally functioned in pathway 1 is suddenly relocated so that it becomes co-expressed with the genes in pathway 2 (and no longer co-expressed with pathway 1), it is possible that its product could begin to act on one of the intermediate molecules in pathway 2. The result would be that the products of pathways 1 and 2 are both changed. Alternative possibilities include those already mentioned in Chapter 1: that there is selection for tighter regulation (*e.g.* if an intermediate molecule in the pathway is toxic), or that there is epistatic selection for tight linkage between interacting alleles.

Chapter 3

SMURF: a web-based tool for identification of fungal secondary metabolite biosynthesis gene clusters

This chapter is based on a manuscript that is in preparation (authors **N. Khaldi**, F.T. Seifuddin, G. Turner, D. Haft, W.C. Nierman, K.H. Wolfe and N.D. Fedorova). I designed the algorithm, and together with Fayaz T. Seifuddin carried out the analyses and wrote the program. I drafted the initial paper and drew the Figures. Kenneth H. Wolfe and Natalie D. Fedorova contributed to the discussion and helped draft the manuscript. Geoff Turner, William C. Nierman, and Natalie D. Fedorova contributed to the planning stage of this project. Daniel Haft developed custom TIGRFAM domain families and together with Fayaz T. Seifuddin performed PFAM and TIGRFAM searches, and contributed to the development of the Perl-CGI web tools.

3.1 Abstract

Fungi produce an impressive array of secondary metabolites (SMs) including mycotoxins, pigments, antibiotics and other biologically active small molecules of great medical, industrial and agricultural importance. The genes responsible for SM biosynthesis, export, and transcriptional regulation are often found in contiguous groups referred to as SM gene clusters, containing groups of neighboring genes that function in the same process but that are mostly unrelated in sequence. Up to now many fungal SM gene clusters remain unannotated in

public databases, so we developed the web-based software SMURF (Secondary Metabolite Unknown Regions Finder, www.jcvi.org/smurf) to search fungal genomes systematically for these pathways. SMURF's predictions are based on PFAM and TIGRFAM domain content as well as on a gene's chromosomal position relative to nearby nonribosomal peptide synthase (NRPS), polyketide synthase (PKS) or prenyl transferase (DMAT) genes, which often encode the first step in SM pathways. In tests, SMURF accurately recovered all the known SM clusters of *Aspergillus fumigatus*, as defined by previous manual annotation and/or by transcription studies in the *laeA* mutants. SMURF detects eight additional potential SM clusters in *A. fumigatus* that were not found by manual annotation. We applied SMURF to catalog the putative SM clusters in 25 publicly available fungal genomes. The analysis reveals the striking variability of SM pathways among fungi at both species and strain levels. It also confirms the correlation between unicellularity and the absence of SMs, and the prominent role of gene duplication in the creation of new pathways. With 200 more fungal genome projects currently underway, SMURF can be used for functional annotation and as a research tool to study the evolution and dynamics of fungal SM pathways.

3.2 Introduction

Secondary metabolites (SMs), also known as natural products, are bioactive molecules produced by many organisms including bacteria, plants and fungi. These compounds are particularly abundant in soil-dwelling filamentous fungi, which exist as multicellular communities competing with each other for nutrients, minerals and water (Keller et al. 2005). Unlike primary metabolites, SMs – as their name suggests – are not essential for survival, growth or reproduction. They can however provide protection against various environmental stresses or fungivory, and facilitate substrate defense or survival inside a eukaryotic host. Scientific appreciation of the importance of fungal SMs grew in the 1940s as the massive impact of penicillin on human health began to be seen. Since then, many other beneficial SM compounds have been discovered including immunosuppressants (e.g., cyclosporin), cholesterol-lowering drugs (lovastatin),

antiviral drugs (equisetin), and anti-tumor drugs (terrequinone A) (for a recent review see (Hoffmeister and Keller 2007)). At the same time, fungi are also known to produce numerous non-beneficial SMs including mycotoxins such as aflatoxin, fumonisin, trichothecene, and zearalenone as well as narcotics such as the hallucinogen molecule LSD.

The first step in biosynthesis of an SM is catalyzed by one of five proteins, which we refer to here as "backbone enzymes": nonribosomal peptide synthases (NRPSs), polyketide synthases (PKSs), hybrid NRPS-PKS enzymes, prenyl transferase (DMATs), and terpene cyclases (TCs). These multidomain enzymes are associated, respectively, with production of the five classes of SMs: nonribosomal peptides, polyketides, hybrids, indole alkaloids, and terpenes (Hoffmeister and Keller 2007). Nonribosomal peptides (e.g., cyclosporin) are peptides that are not formed by ribosomal translation of mRNAs, but by untemplated chemical condensation of amino acid monomers by NRPS enzymes (Marahiel et al. 1997; Stachelhaus et al. 1998; Finking and Marahiel 2004). Notably, NRPS-like enzymes without a condensation domain can also be involved in production of SMs such as fusarin, a mycotoxin produced by some Fusarium species (Song et al. 2004). Likewise, fungal polyketides (e.g., aflatoxin) are synthesized by multidomain PKSs from acyl-coenzyme A (Hopwood and Sherman 1990; Hopwood 1997). Hybrid PKS-NRPS enzymes, only recently identified in fungi (Kroken et al. 2003; Bohnert et al. 2004; Song et al. 2004), are involved in biosynthesis of hybrid metabolites such as equisetin (Burke et al. 2005; Sims et al. 2005). Indole alkaloids (e.g., ergotamine) are composed primarily of dimethylallyl diphosphate units, linked to tryptophan or other amino acids (Tudzynski et al. 1999). They include ergopeptides (Haarmann et al. 2006), and fumigaclavines (Schardl et al. 2006). Terpenes, which are composed of isoprene units, are not considered further in our analysis because terpene cyclases are highly variable in sequence and difficult to detect by bioinformatic methods (Townsend 1997; Keller et al. 2005). The two most studied terpenes are trichothecene and gibberellin.

Intermediate products formed by the backbone enzymes can undergo further modifications such as condensation, methylation and oxygenation. These steps are

often catalyzed by enzymes belonging to a few domain superfamilies such as cytochrome P450 oxidoreductases, methyltransferases and enoyl-reductases. The final product is then secreted out of the cell by a transporter, typically of the MFS (major facilitator superfamily) or ABC (ATP-binding cassette) transporter families. The whole process is often regulated by a Zn_2Cys_6 transcription factor.

An unusual and striking feature shared by most fungal SM synthesis pathways is that many of the genes coding for components of these pathways are found in close physical proximity to their "backbone" genes, forming SM gene clusters (Keller and Hohn 1997). Some of the known SM gene clusters are very large, such as the 25-gene cluster for sterigmatocystin synthesis in *Aspergillus nidulans* (Brown et al. 1996), which is – with the exception of the major histocompatibility locus in mammals – the largest known physical cluster of genes that are co-functional (but unrelated in sequence) in any eukaryote. Although some examples are known of SM biosynthesis pathways whose genes are not clustered, the majority of fungal SM pathways that have been described to date involve clustered genes and it is clear that clustering is the primary mode of organization of fungal SM genes (Keller et al. 2005).

Despite the medical and agricultural importance of SMs and the recent availability of many complete fungal genome sequences, relatively few SM pathways have been characterized at the molecular level so far. In previous analyses of sequenced fungal genomes, putative SM gene clusters were identified by ad hoc strategies based on BLAST searches; for example, 22 clusters have been proposed in the *A. fumigatus* genome (Nierman et al. 2005; Perrin et al. 2007). As well as being time-consuming, manual annotation runs the risk that inconsistent approaches might be used for different genomes or even for different clusters within the same genome. In many cases it can be difficult to decide exactly where a cluster begins and ends, particularly if transcription data is not available. Here, we developed a bioinformatics approach for finding SM gene clusters in sequenced genomes. We used three hallmarks of SM biosynthetic pathways (the presence of backbone genes; clustering; and characteristic protein domain content) to systematically predict the SM gene clusters that are present in 25 sequenced fungal genomes. We show SM gene enrichment in the genus Aspergillus, the absence of SM gene

clusters in unicellular fungi, and unexpected variability of the SM gene cluster content at both species and strain levels. Our results are consistent with the view that SM profiles can be used as means of differentiating species and strains in filamentous fungi, and that gene duplication plays an essential role in the creation and expansion of the SM repertoires of fungi.

3.3 Methods

3.3.1 Detecting backbone enzymes.

The HMMER program (http://hmmer.janelia.org/) was used to search for conserved Pfam domains of backbone enzymes in the protein set of each sequenced species. Trusted threshold cutoffs (predefined in HMMER) were used for each HMM search. NRPS enzymes were identified with at least a minimal composition of one module composed of an amino acid adenylation domain (A), a thiolation domain (PCP) and a condensation domain (C). PKS enzymes were identified with at least one acyl transferase domain (AT), a beta-ketoacyl synthase C-terminal domain (BKS-C), and a beta-ketoacyl synthase N-terminal domain (BKS-N). Hybrid PKS-NRPS enzymes were identified with at least all of the PKS or NRPS domains above. prenyl transferase enzymes were identified with at least one prenyl transferase domain, an aromatic prenyltransferase domain, a prenyl transferase type domain, and a tryptophan dimethylallyl transferase domain. NRPS-like enzymes were identified with a combination of at least two domains from any of those in the NRPS enzyme module; or a combination of an A domain and a NAD_binding_4 domain; or a combination of an A domain and short chain dehydrogenase domain. PKS-like enzymes were identified with a combination of at least two domains from any of those in the PKS enzyme module.

One of the problems we encountered involved distinguishing true backbone enzymes from enzymes with similar domain composition, which are involved in primary metabolism. To this end, for PKS-like genes we required the scores for AT, BKS-C and BKS-N domains to be below trusted HMM cut-off. To eliminate false positives such as alpha-aminoadipate reductase (An et al. 2003) among

NRPSs, we required the score of the C-terminal domain of L-aminoadipate-semialdehyde dehydrogenase alpha subunit to be above trusted HMM cut-off.

3.3.2 Detecting SM clusters.

We first defined 27 protein domains commonly found in SM clusters (Table 1). We used HMMER to search for these domains in the protein set of each species, applying HMMER's trusted threshold cutoff scores for each domain.

The SMURF algorithm then evaluates each backbone enzyme gene to test whether it forms part of a cluster. The process consists of two major parts: (i) finding the preliminary extent of the cluster, and (ii) trimming the boundaries of the cluster. For the first part, we begin at the backbone enzyme gene and start to walk rightwards. Each gene that we meet on this walk is classified as either SM domain-positive (if it contains any of the 27 domains in Table 1) or SM domain-negative (otherwise). We continue walking rightwards until we reach a stop signal. A stop signal is defined as either an intergenic spacer distance that is larger than the limit d, or a cumulative number of SM domain-negative genes between the backbone gene and the current position that is larger than the limit y. The last gene on the rightwards walk before the stop signal was encountered is given the label alpha. We then carry out an identical walk leftwards, until a stop signal is encountered defining a left-limit gene beta. The interval between alpha and beta is the preliminary extent of the cluster.

In the second part of the process we trim the boundaries of the cluster. Trimming is likely to be necessary for two reasons. The total number of SM domain-negative genes in the interval between alpha and beta may be greater than y (it could be as large as $2y$). Also, if the genes alpha or beta are neither SM domain-positive nor backbones, we will want to reduce the limits of the cluster to the outermost genes that are. To trim the cluster, we consider every possible interval within it that runs from one SM domain-positive (or backbone) gene to another. For each of these intervals, if its total content of SM domain-negative genes is less than y, we calculate its score. An interval's score is defined as the total number of

SM domain-positive genes and backbone genes within it. The interval with the highest score is output as the final cluster; if two intervals have the same score then we arbitrarily output one of them.

3.4 Results

3.4.1 Overview of SMURF

SMURF operates by detecting groups of nearby genes that contain particular protein domains. To choose the set of domains for SMURF, we considered all known clusters that have been identified either by standard genetic methods in any filamentous fungus, or by transcriptional profiling of *laeA* regulatory mutants in *Aspergillus fumigatus* strain Af293 (Perrin et al. 2007b). We thus identified 27 protein domains that are commonly found in SM clusters; we refer to these as SM domains (Table 3.1). For most of the domains one or more corresponding domain models already existed in the PFAM database (Mistry et al. 2007), and we built a new model for one other domain (N-methyltransferase).

Once the required protein sequence data are uploaded, SMURF uses hidden Markov model (HMM) searches to detect backbone genes in each proteome. After this step, a window of ±20 genes on each side of the backbone gene is scanned for the 27 SM domains using HMM searches. Genes in the window are tagged as 0 or 1 depending on whether they contain the searched domain. Then we delimit the boundaries of the cluster by applying an algorithm that optimizes the score of the putative cluster when the presence of genes lacking SM domains within the cluster is penalized (see *Methods* 3.3). In addition to the presence of SM domains, SMURF also considers the intergenic distances between genes because it is known that SM gene clusters tend to have exceptionally short intergenic distances (Keller et al. 2005). In some instances SMURF can find the same SM cluster twice, if it contains more than one backbone gene, in which case the two putative clusters are merged.

3.4.2 Training set

The SMURF algorithm requires two key parameters: *d*, the maximum intergenic distance (in base pairs) permitted between two adjacent genes in the same cluster;

Table 3.1 The 27 PFAM domains searched for by SMURF, and their distribution in the *A. fumigatus* Af293 genome.

PFAM number	Domain name	Number of domains in genome	Number of domains in clusters	Fraction in clusters
PF01244	Renal dipeptidase	1	1	1.00
PF02458	Transferase family	6	4	0.67
PF00891	O-methyltransferase	10	5	0.50
PF01227	GTP cyclohydrolase I	2	1	0.50
PF05721	Phytanoyl-CoA dioxygenase (PhyH)	7	3	0.43
PF00753	Metallo-beta-lactamase domain	9	3	0.33
PF00043	Glutathione S-transferase, C-terminal domain	10	3	0.30
PF03171	Oxidoreductase, 2OG-Fe(II) oxygenase family	14	4	0.29
PF00067	Cytochrome P450	69	18	0.26
PF00557	Peptidase, M24 family	11	2	0.18
PF00106	Oxidoreductase, short chain dehydrogenase/reductase family	122	18	0.15
PF00664	ABC transporter transmembrane region	22	3	0.14
PF01494	FAD binding domain	39	5	0.13
PF01565	FAD binding domain	32	4	0.13
PF02798	Glutathione S-transferase, N-terminal domain	16	2	0.13
PF00005	ABC transporter, ATP-binding protein	51	5	0.10
PF00155	Aminotransferase, classes I and II	24	2	0.08
PF00107	Oxidoreductase, zinc-binding dehydrogenase family	49	4	0.08
PF00070	Pyridine nucleotide-disulphide oxidoreductase	14	1	0.07
PF00172	Fungal Zn(2)-Cys(6) binuclear cluster domain	208	13	0.06
PF07992	Pyridine nucleotide-disulphide oxidoreductase	19	1	0.05
PF07690	Transporter, major facilitator family	286	11	0.04
PF00583	Acetyltransferase, GNAT family	40	1	0.03
PF04082	Fungal specific transcription factor domain	126	3	0.02
PF00083	Transporter, major facilitator superfamily	94	2	0.02
PF00324	Amino acid permease (subfamily of MSF transporter)	105	0	0
TIGR03439*	Probable methyltransferase domain, EasF family	7	2	0.29

Note- * An HMM model for the N-methyltransferase domain was created in this study.

and *y*, the maximum number of genes that do not posses one of the 27 SM domains that is permitted to exist within a cluster. By a trial-and-error process we identified the parameters *d* = 3814 bp and *y* = 10 genes as optimal for the recovery of the genetically known clusters in the *A. fumigatus* genome. We optimized the parameters *d* and *y* by considering data from the following 15 SM clusters that have been experimentally characterized: *A. fumigatus* clusters *Pes1* (Reeves et al. 2006), siderophore (Schwecke et al. 2006), ETP (epipolythiodioxopiperazine`; Gardiner et al. 2005b), fumitremorgin (Maiya et al. 2006), gliotoxin (Gardiner and Howlett 2005), ergot alkaloid (Coyle and Panaccione 2005), fumigaclavine (Unsold and Li 2006), melanin (Tsai et al. 1999), and pseurotin (Maiya et al. 2007); *A. nidulans* sterigmatocystin (Brown et al. 1996); *A. flavus* aflatoxin (Yu et al. 1995) and aflatrem (Zhang et al. 2004); *Gibberella zeae* aurofusarin (Malz et al. 2005), and zearalenone (Kim et al. 2005); *G. verticillioides* fumonisin (Proctor et al. 2003).

After parameter optimization with this training set, SMURF was able to correctly recover all of the above SM clusters from their respective genome sequences. SMURF tends to predict clusters that are somewhat larger than those annotated manually: in the training set the mean number of extra genes over-predicted is seven genes per cluster. The mean was largely affected by the overprediction of 20 genes in the pseurotin cluster of *A. fumigatus* as compared to the manual annotation (Table 3.2). However, this overprediction is due to the fact that this region of the genome was manually annotated as containing two clusters (the pseurotin cluster AFUA_8G00530-AFUA00580, and a neighboring cluster AFUA_8G00290–AFUA_8G00520), which were merged by SMURF into a larger single cluster because there are no intervening genes between them.

Only in two (out of eight) cases did SMURF under-predict the boundaries of a cluster whose boundaries have been experimentally identified in *A. fumigatus*. One case of under-prediction occurred in the Pes1 cluster, which was manually annotated as containing only two genes. SMURF omitted one of these genes because the spacer between them was unusually long, but SMURF also identified additional SM domain-containing genes in the neighborhood and reported a nine-gene cluster in total. The second case of under-prediction occurred in the

fumitremorgin cluster, for which SMURF correctly predicted 18 genes but misses the last one (AFUA_8G00290) because this particular gene is hypothetical and codes for a protein of unknown metabolic function.

One of the key challenges in finding SM cluster boundaries is the large number of hypothetical genes found in them. This is prevalent in many well-characterized SM clusters such as sterigmatocystin, in this case the hypothetical genes have been found to be essential (McDonald et al. 2005). The parameters in SMURF were chosen to reduce under-prediction caused by many of the hypothetical genes that have not yet been characterized. This explains SMURF's over-prediction. Indeed, SMURF is confronted with the dilemma of balancing both under and over-predictions; we favored the later, which as a result over-predicts the boundaries of a potential SM cluster by a mean of seven genes. Over-prediction can be addressed in the future once more secondary metabolite gene domains are characterized.

3.4.3 Application to *Aspergillus fumigatus*

The SM cluster content of *A. fumigatus* strain Af293 has previously been annotated carefully by hand (Nierman et al. 2005; Perrin et al. 2007), and 22 SM clusters were found by this manual approach. Of these, 13 have been confirmed experimentally because they are regulated by the master SM transcription factor *laeA* (Perrin et al. 2007b). Eight of the 22 *A. fumigatus* SM clusters have known functions and were also in our training set.

SMURF successfully found all 22 manually annotated clusters in *A. fumigatus* (Table 3.2), including all eight that have been detected experimentally. As mentioned above, SMURF merged the pseurotin cluster with an immediately neighboring SM gene cluster whose product is unknown. Comparison of the boundaries of the 14 manually identified SM clusters in *A. fumigatus* (which are not experimentally validated yet, and thus constitute a less reliable set for testing SMURF, especially regarding the delimitation of cluster boundaries) shows that SMURF under-predicted five clusters (by 1-4 genes each, mostly again due to

unusually large intergenic distances), with a mean of one gene under-predicted per cluster. In this set of 14 clusters SMURF again slightly over-predict the extent of the clusters, with a mean of seven over-predicted genes per cluster, the same as in the experimentally validated clusters.

More surprisingly, SMURF also identified a further eight clusters in *A. fumigatus* that were not identified by the manual annotation (Table 3.2). We further examined these eight clusters by eye and conclude that five of them contained essential genes implicated in secondary metabolism synthesis, and thus are potential SM clusters. We looked at the *laeA* profile for these eight clusters and we did not find any where all the genes were consecutively induced (with no gaps). This does not argue against the validly of these clusters. Many of the manually annotated clusters were also not up-regulated in the *laeA* deletion strain, but contained many essential genes implicated in secondary metabolism synthesis. Examining the remaining three clusters proposed by SMURF shows that they are missing necessary domains, and thus are not chemically plausible to give rise to SMs. The small number of these types of clusters with uncommon combinations show that their occurrence is not random. They maybe the remains of ancestral clusters that are no longer functional, or they might be components of SM pathways whose genes are partly scattered. In both cases, an evolutionary study of the clusters might shed light on how SM clusters evolve.

Table 3.2 Comparison of *Aspergillus fumigatus* SM clusters annotated automatically using SMURF to those annotated manually.

SMURF annotation			Test set clusters				
Backbone†	Cluster Start	Cluster Stop	Found manually?	Cluster Start	Cluster Stop	Difference**	Product
			SM clusters identified experimentally				
1G10380	1G10270	1G10380	yes	1G10380	1G10390	-1, +8	Pes1
2G17600	2G17460	2G17700	yes	2G17530	2G17600	+2	Melanin
2G17990[a]	2G17930	2G18070	yes	2G17960	2G18060	0	Ergot
3G03350b	3G03270	3G03550	yes	3G03350	3G03470	+15	Siderophore
3G12920[c]	3G12750	3G13060	yes	3G12890	3G12960	+10	ETP
6G09610[d]	6G09570	6G09745	yes	6G09590	6G09740	+1	Gliotoxin
8G00210[e]	8G00100	8G00280	yes	8G00100	8G00290	-1,0	Fumitremorgin
8G00540[f]	8G00370	8G00620	yes	8G00530	8G00580	+20	Pseurotin
			*SM clusters found by manual annotation**				
1G17740	1G17710	1G17750	yes	1G17710	1G17740	0	
3G01410	3G01400	3G01580	yes	3G01400	3G01420	+13	
3G02570	3G02550	3G02650	yes	3G02560	3G02630	-2	
3G13730	3G13570	3G13750	yes	3G13710	3G13730	+11	
3G14700	3G14665	3G14850	yes	3G14690	3G14720	+11	
3G15270	3G15150	3G15290	yes	3G15250	3G15310	-1	
4G00210	4G00195	4G00350	yes	4G00150	4G00235	-3, +3	
4G14560	4G14420	4G14735	yes	4G14520	4G14590	+12	
5G12730	5G12620	5G12850	yes	5G12720	5G12740	+10	
6G03480	6G03430	6G03630	yes	6G03400	6G03490	-1, +12	
6G12050[g]	6G11930	6G12190	yes	6G12040	6G12080	+9	
6G13930	6G13790	6G14090	yes	6G13920	6G14000	+12	
7G00160[h]	7G00120	7G00200	yes	7G00120	7G00180	-2	
8G00370[i]	8G00342	8G00570	yes	8G00300	8G00520	-4, +4	
			SM clusters not detected manually				
1G17200[j]	1G17070	1G17290	no				
2G01290[k]	2G01190	2G01410	no				
3G02530[l]	3G02450	3G02540	no				
3G02670[m]	3G02670	3G02760	no				
5G10120[n]	5G09970	5G10270	no				
6G08560	6G08520	6G08660	no				
8G01640	8G01640	8G01640	no				
8G02350	8G02490	8G02350	no				

Note- * Manual annotation is based on Perrin et al. (2007a) with some additional curation.
** Difference between the cluster content predicted by SMURF and that predicted manually. Negative numbers count genes found manually but missed by SMURF. Positive numbers count genes included in SMURF's clusters but not in the manual clusters.
† The prefix AFUA_ has been omitted from all gene names to save space.
[a] The same cluster was also detected by SMURF with backbone AFUA_2G18040.
[b] The same cluster was also detected by SMURF with backbone AFUA_3G03420.
[c] The same cluster was also detected by SMURF with backbone AFUA_3G12930.
[d] The same cluster was also detected by SMURF with backbone AFUA_6G09660.
[e] The same cluster was also detected by SMURF with backbones AFUA_8G00170 and AFUA_8G00250.
[f] The same cluster was also detected by SMURF with backbone AFUA_8G00540.
[g] The same cluster was also detected by SMURF with backbone AFUA_6G12080.
[h] The same cluster was also detected by SMURF with backbone AFUA_7G00170.
[i] The same cluster was also detected by SMURF with backbone AFUA_8G00540.
[j], [k], [l], [m], and [n] are the extra 5 potential clusters determined by SMURF and not detected by manual annotation.

3.4.4 Uneven taxonomic distribution of backbone enzymes

We systematically searched the genome sequences of 25 fungal species (22 Ascomycota and three Basidiomycota; Table 3.3) for the presence of four types of backbone genes: NRPS, PKS, PKS-NRPS hybrids, and prenyl transferase. As expected, the search revealed that the numbers of putative backbone enzymes varies greatly (from 0-61) from one fungal taxon to another (Figure 3.1 A). There are no backbone genes in two of the three unicellular species we considered, the ascomycete yeast *Saccharomyces cerevisiae*, and the basidiomycete yeast *Cryptococcus neoformans*. Similarly, there is only one backbone gene (an NRPS; (Schwecke et al. 2006)) in the genome of a third unicellular fungus, the archiascomycete yeast *Schizosaccharomyces pombe*.

We also catalogued NRPS-like and PKS-like enzymes because we suspected that many SM cluster genes could have unusual domain structures; for example the fusarin C cluster of *Fusarium* species includes a gene with a typical NRPS condensation domain but nothing else (Song et al. 2004). We find that the numbers of NRPS-like and PKS-like genes fluctuate in correlation with their counterparts the canonical NRPS and PKS genes (Figure 3.1 A).

Figure 3.1 A also shows larger numbers of backbone genes in Pezizomycotina, especially in Eurotiomycetes, as compared to Basidiomycota and Sordariomycota. However, within the Eurotiomycetes, there are notably fewer backbone genes in the genomes of the human pathogens *Coccidioides immitis* and *C. posadasii* than in the section Aspergillus. This difference is more likely to be due to the phylogenetic distance between *Coccidioides* and *Aspergillus* than to lifestyle differences, because the human pathogen *A. terreus* retains a large arsenal of 59 backbone genes.

Among the Pezizomycotina, *Neurospora crassa* has a significantly reduced number of backbone genes (10), even when compared to *Fusarium oxysporum* which has the second lowest number of backbone genes in the Pezizomycotina ($P < 10^{-16}$, Chi-square test). This difference is presumably attributable to the presence

of a very active repeat-induced point mutation (RIP) process in *N. crassa*, which has dramatically reduced the rate of formation of new gene duplications in that species (Galagan et al. 2003).

PKSs and NRPSs are found in significantly higher numbers than prenyl transferase and hybrid enzymes in almost all species. We also observed that the number of backbone genes is significantly higher in the Aspergilli than in Sordariomycetes ($P = 0.001$, Wilcoxon test). This difference is due to increases in the numbers of NRPS ($P = 7 \times 10^{-4}$), PKS ($P = 0.001$), and prenyl transferase ($P = 0.002$) enzymes in Aspergilli, but not hybrid enzymes ($P = 0.9$).

We found no significant correlation between the number of enzymes and the genome size. There is no significant difference between the numbers of enzymes in pathogens versus non-pathogens, nor between human pathogens versus plant pathogens. We did however observe that the percentage of genes in the genome that are backbone genes is lower in plant pathogens versus fungi with all any other lifestyle ($P = 0.001$ or $P= 0.03$ depending on whether *N. crassa* is excluded or included, respectively).

Table 3.3 Complete fungal genomes used in the search for SM clusters.

Species	Strain	Source	Sequencing center	Taxonomy	Anamorph
Saccharomyces cerevisiae	S288c	www.yeastgenome.org	Intl. Consortium	Ascomycota; Saccharomycotina; Saccharomycetes	
Schizosaccharomyces pombe	972h⁻	www.sanger.ac.uk	Intl. Consortium	Ascomycota; Taphrinomycota; Schizosaccharomycetes	
Phaeosphaeria nodorum	SN15	www.broad.mit.edu	Broad Institute	Ascomycota; Pezizomycotina; Dothideomycetes	Stagonospora nodorum
Aspergillus clavatus	NRRL 1	www.ncbi.nlm.nih.gov	TIGR/JCVI	Ascomycota; Pezizomycotina; Eurotiomycetes	
Aspergillus flavus	NRRL3357	www.aspergillusflavus.org	TIGR/JCVI/NCSU	Ascomycota; Pezizomycotina; Eurotiomycetes	
Aspergillus fumigatus	Af293	www.ncbi.nlm.nih.gov	TIGR/JCVI	Ascomycota; Pezizomycotina; Eurotiomycetes	
"	CEA10	www.ncbi.nlm.nih.gov	TIGR/Celera	Ascomycota; Pezizomycotina; Eurotiomycetes	
Aspergillus niger	ATCC 1015	img.jgi.doe.gov	JGI	Ascomycota; Pezizomycotina; Eurotiomycetes	
"	CBS 513.88	www.ncbi.nlm.nih.gov	DSM	Ascomycota; Pezizomycotina; Eurotiomycetes	
Aspergillus oryzae	RIB40	www.bio.nite.go.jp	NITE	Ascomycota; Pezizomycotina; Eurotiomycetes	
Aspergillus terreus	NIH2624	www.broad.mit.edu	Broad Institute	Ascomycota; Pezizomycotina; Eurotiomycetes	
Coccidioides immitis	H538.4	www.broad.mit.edu	Broad Institute	Ascomycota; Pezizomycotina; Eurotiomycetes	
"	RMSCC 2394	www.broad.mit.edu	Broad Institute	Ascomycota; Pezizomycotina; Eurotiomycetes	
Coccidioides posadasii	C735	www.broad.mit.edu	TIGR/Broad	Ascomycota; Pezizomycotina; Eurotiomycetes	
Emericella nidulans	FGSC A4	www.ncbi.nlm.nih.gov	Broad Institute	Ascomycota; Pezizomycotina; Eurotiomycetes	*Aspergillus nidulans*
Neosartorya fischeri	NRRL 181	www.ncbi.nlm.nih.gov	TIGR/JCVI	Ascomycota; Pezizomycotina; Eurotiomycetes	*Aspergillus fischerianus*
Uncinocarpus reesii	1704	www.broad.mit.edu	Broad Institute	Ascomycota; Pezizomycotina; Eurotiomycetes	*Malbranchea reesii*
Sclerotinia sclerotiorum	ATCC18683	www.broad.mit.edu	Broad Institute	Ascomycota; Pezizomycotina; Leotiomycetes	
Fusarium oxysporum	NRRL 34936	www.broad.mit.edu	Broad Institute	Ascomycota; Pezizomycotina; Sordariomycetes	
Gibberella moniliformis	NRRL 20956	www.broad.mit.edu	Broad Institute	Ascomycota; Pezizomycotina; Sordariomycetes	*Fusarium verticillioides*
Gibberella zeae	NRRL 31084	www.broad.mit.edu	Broad Institute	Ascomycota; Pezizomycotina; Sordariomycetes	*Fusarium graminearum*
Magnaporthe grisea	70-15	www.ncbi.nlm.nih.gov	Broad Institute	Ascomycota; Pezizomycotina; Sordariomycetes	*Pyricularia oryzae*
Neurospora crassa	OR74A	www.broad.mit.edu	Broad Institute	Ascomycota; Pezizomycotina; Sordariomycetes	
Chaetomium globosum	CBS 148.51	www.broad.mit.edu	Broad Institute	Ascomycota; Pezizomycotina; Sordariomycetes	
Hypocrea jecorina	QM6a	img.jgi.doe.gov	JGI	Ascomycota; Pezizomycotina; Sordariomycetes	*Trichoderma reesei*
Phanerochaete chrysosporium	RP-78	img.jgi.doe.gov	JGI	Basidiomycota; Agaricomycotina; Agaricomycetes	
Cryptococcus neoformans	R265	www.ncbi.nlm.nih.gov	Broad Institute	Basidiomycota; Agaricomycotina; Tremellomycetes	*Filobasidiella neoformans*
Ustilago maydis	521	www.ncbi.nlm.nih.gov	Broad Institute	Basidiomycota; Ustilaginomycotina; Ustilaginomycetes	

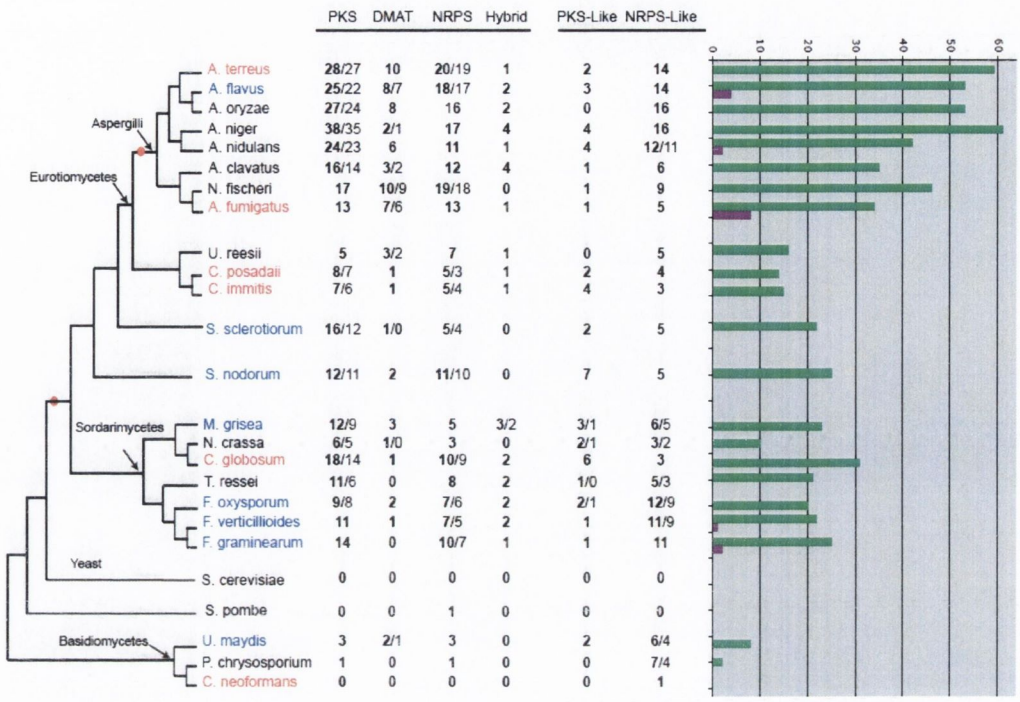


Figure 3.1 Numbers of backbone genes and SM clusters in the 25 fungal genomes we analyzed. (A) Phylogenetic relationships (Fitzpatrick et al. 2006) and numbers of genes of each domain type. Each column contains two numbers; the first (in bold) is the number of backbone genes, and the second is the number of putative SM clusters containing this backbone. Species named in red are human pathogens, blue are plant pathogens, and black are non-pathogenic fungi. Red bullets mark two internal branches on which enrichment in backbone genes occurred. (B) Green bars show the total numbers of potential SM clusters (excluding SM clusters only containing PKS-like and NRPS-like) identified by SMURF in each species. Purple bars show the numbers of SM clusters that have been identified experimentally.

3.4.5 Species specificity of SM clusters

The large numbers of putative SM gene clusters identified by SMURF (Figure 3.1 B) emphasizes the huge diversity of the repertoires of metabolites synthesized by fungal species. To what extent are these metabolites species-specific? We cannot answer this question directly, but we can obtain an approximate answer by investigating the extent to which the SM clusters are species-specific.

We further analyzed the three closely related genomes *A. fumigatus* Af293, *A. clavatus* and *Neosartorya fischeri*. For each putative SM cluster predicted by SMURF in each of these species, we BLAST searched the genes against the other two genomes. We scored an SM cluster from one species as being conserved in a

second species if at least 80% of its genes were clustered in the second species. The results (Figure 3.2) show that only five SM clusters are common to all three genomes. This core set of includes clusters that are present in many other fungi, such as Pes1 (Reeves et al. 2006), siderophore (Schwecke et al. 2006), and melanin (Tsai et al. 1999). The majority of SM clusters are specific to only one of the three species, and the functions of most of these clusters are unknown.

We were also curious to analyze the extent of within-species polymorphism for the presence or absence of SM clusters. To do this, we applied the same search strategy outlined above to the sequenced genomes of two strains of *A. fumigatus* Af293 and CEA10. This search showed that two SM clusters present in Af293 are absent from CEA10. We then searched for orthologs of these genes in the other sequenced *Aspergillus* species. One five-gene cluster present in Af293 (AFUA_1G17710–AFUA1G17740) but not CEA10 has an orthologous cluster in *A. clavatus* (ACLA_098870–ACLA_098920) but not in any other *Aspergillus* species shown in Figure 3.1 . The other Af293-specific cluster has no orthologs in any other species. These results indicate dynamic manner in which SM gene cluster repertoire can turn over, even within a single species.

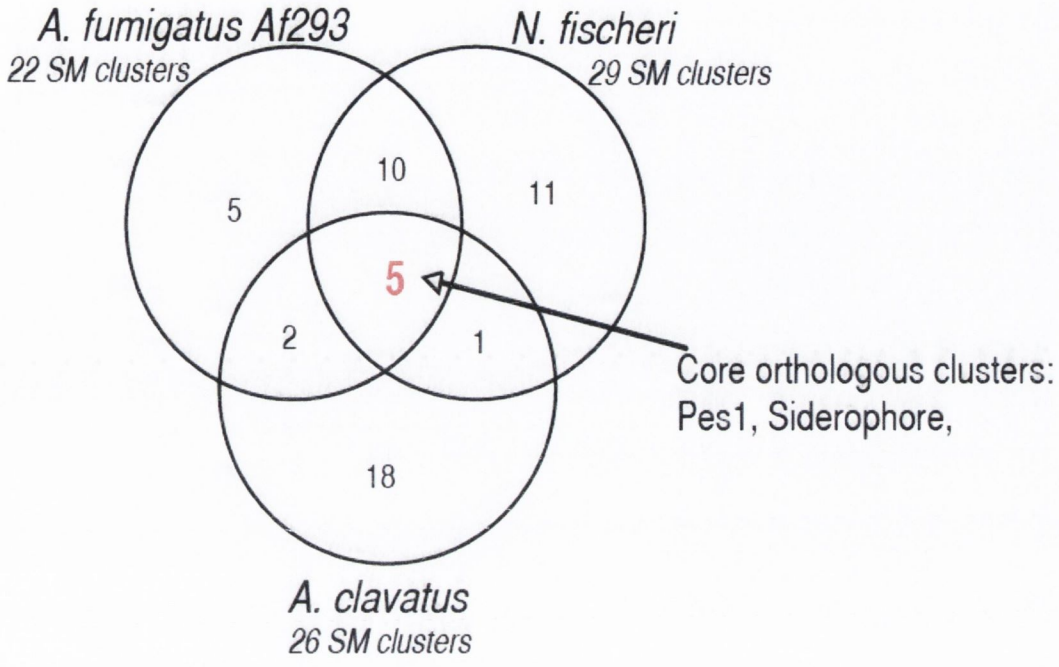


Figure 3.2 Venn diagram showing numbers of predicted SM clusters that are shared among *A. fumigatus* Af293, *A. clavatus* and *N. fischeri*.

3.5 Discussion

Comparison of SM clusters identified by genetic studies to those predicted by SMURF demonstrates that the algorithm has high sensitivity and specificity (Table 3.2). Many SM clusters, including sterigmatocystin, lovastatin, pigments, gliotoxin and penicillin, are regulated by the global regulator of secondary metabolism, LaeA (Bok and Keller 2004; Keller et al. 2006). SMURF identified eight additional clusters that had not been found in previous manual annotations of SM clusters in *A. fumigatus*. Of these eight, five have organizations and contents typical of SM clusters, which indicates that these are likely to be functional SM clusters and demonstrates the value of automated approaches to searching. While SMURF estimates the approximate boundaries of its SM clusters based on sequence properties, expression studies in *laeA* mutants can be used to

delineate the boundaries more precisely, as recently shown for terrequinone A, an anti-tumor agent produced by *A. nidulans* (Bouhired et al. 2007).

Our results suggest that the numbers of potential SM produced by fungi is even higher than previously anticipated, with nonribosomal peptides and polyketides being the most abundant molecules. Among the taxa studied by genome sequencing, the Aspergilli appear to have both the largest numbers and the greatest intraspecies variability of these metabolites. The genome of *A. niger* alone encodes 61 backbone enzymes and 58 clusters. Further comparison reveals that SM pathways are very unevenly distributed among fungal taxa suggesting an evolutionary pattern of rapid pathway gain and loss.

It is tempting to speculate about the selective pressures that can cause this amazing chemical diversity and its uneven taxonomic distribution. Previously the diversity of fungal natural products has been attributed to diversifying selection driven by a "chemical arms race" with predators and other soil inhabitants (Czaran et al. 2002). Our results indicate a correlation between the presence of SM pathways and the filamentous growth form among fungal taxa. The species phylogeny shows that numerous SM backbone genes were lost, on three separate occasions, on branches leading to species with unicellular growth habits (*S. cerevisiae*, *Sch. pombe*, and *C. neoformans*). We can infer that losses of secondary metabolism genes coincided with transitions to a unicellular lifestyle during the evolution of each of these yeast lineages (Figure 1), because the ancestor of all fungi was multicellular (Liu and Hall 2004).

Indeed, the lifestyles of filamentous fungi are in some ways very similar to those of plants. For the most part of their life cycle, they are immobile and have to compete for the same resources such water, nutrients, minerals with other soil inhabitants. They also need to avoid predation by insects, amoebae and protozoa. As a result, filamentous fungi have to maintain formidable arrays of SMs that provide fitness benefits under a variety of circumstances. In contrast, unicellular fungi are more mobile and thus have the option of "flight" instead of "fight". They do not need SMs as defense mechanisms. Interestingly the core clusters that we identified as orthologous in the Aspergilli (Figure 3.2) are involved in

biosynthesis of the Pes1 peptide, siderophore and spore pigment, all of which have been linked to protection against oxidative stress (Eisendle et al. 2003; Han et al. 2004; Reeves et al. 2006; Schrettl et al. 2007). Perhaps the more divergent SM clusters are those involved in chemical arms races with other organisms including competitors, predators and hosts. We found fewer backbone genes in plant pathogens than in other fungi. A reason for this might be that plant pathogens tend to be specific to particular hosts and thus face only a finite number of conditions. In contrast, animal pathogens and non-pathogenic fungi may be more exposed to conditions that can differ dramatically, and in consequence their arsenal is more diverse.

Cross-species comparison of SM clusters shows that very few of them are shared even among very closely related fungi (Figure 3.2). This is consistent with the view that most fungal SMs are lineage-specific. It has been suggested that secondary metabolism can be used as species markers (Frisvad and Samson 2004; Frisvad et al. 2005). For example, Frisvad and Samson showed that the secondary metabolite asteltoxin can be a good chemotaxonomic marker for *Emericella variecolor* (Frisvad and Samson 2004). Our results suggest that, depending on the metabolite, SM profiles could be used as diagnostic markers at either an inter-species or an inter-strain level.

Filamentous ascomycetes and Aspergilli in particular provide a pharmacopoeia of mycotoxins, antibiotics and other natural products and genome sequencing projects greatly facilitate the discovery of new compounds. Thus within months after completion of the *A. fumigatus* genome and manual identification of its putative SM clusters (Nierman et al. 2005), several secondary metabolite clusters were characterized at the molecular level including gliotoxin (Gardiner and Howlett 2005), ergot alkaloids (Coyle and Panaccione 2005), fumigaclavines (Unsold and Li 2006), fumitremorgin (Maiya et al. 2006), and hydroxamate siderophore biosynthesis clusters (Schwecke et al. 2006). As a web-based tool or a stand-alone application, SMURF can further accelerate the discovery of new pathways in hundreds of other fungal genomes currently being sequenced (GOLD database, www.genomesonline.org), These SM pathways can be characterized

experimentally and linked to specific products which may produce compounds of medical, agricultural and industrial importance.

SMURF has considerable potential for future growth and development. For example, the accuracy of SM cluster boundary predictions by SMURF could be increased by the addition of more protein domains. In addition, SMURF could be fine-tuned to search only for particular pathways, by restricting it to allow only particular combination of genes. Fine-tuning of this type could also further reduce the false-positive rate (3 of the 30 clusters that SMURF predicted in the *A. fumigatus* genome appear to be false positives). Finally, we note that the density of genes is a key factor in identifying SM clusters, but the density of genes within SM clusters varies in different subgroups and SMURF does not differentiate these subgroups. Providing SMURF with details of the species subgroup could further improve accuracy.

Chapter 4

Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi

4.1 Related publication

The research described in this chapter has been published (**N. Khaldi***, J. Collemare*, M.-H. Lebrun, K. H. Wolfe, *Genome Biology* 9:R18, 2008). *These authors contributed equally to this work. Jérôme Collemare and Marc-Henri Lebrun isolated the *M. grisea ACE1* cluster. The author and Jérôme Collemare conducted genome searches and phylogenetic analyses. Kenneth H. Wolfe drew the figures. All authors contributed to writing the manuscript.

4.2 Abstract

Filamentous fungi synthesize many secondary metabolites and are rich in genes encoding proteins involved in their biosynthesis. Genes from the same pathway are often clustered and co-expressed in particular conditions. Such secondary metabolism gene clusters evolve rapidly through multiple rearrangements, duplications and losses. It has long been suspected that clusters can be transferred horizontally between species, but few concrete examples have been described so far. In the rice blast fungus *Magnaporthe grisea*, the avirulence gene *ACE1* that codes for a hybrid PKS-NRPS belongs to a cluster of 15 genes involved in secondary metabolism. Additional related clusters were detected in the ascomycetes *Chaetomium globosum*, *Stagonospora nodorum* and *Aspergillus clavatus*. Gene-by-gene phylogenetic analysis showed that in *C. globosum* and *M.*

grisea, the evolution of these *ACE1*-like clusters is characterized by successive complex duplication events including tandem duplication within the *M. grisea* cluster. The phylogenetic trees also present evidence that at least five of the six genes in the homologous *ACE1* gene cluster in *A. clavatus* originated by horizontal transfer from a donor closely related to *M. grisea.*

The *ACE1* cluster originally identified in *M. grisea* is shared by only few fungal species. Its sporadic distribution within euascomycetes is mainly explained by multiple events of duplication and losses. However, because *A. clavatus* contains an *ACE1* cluster of only six genes, we propose that horizontal transfer from a relative of *M. grisea* into an ancestor of *A. clavatus* provides a much simpler explanation of the observed data than the alternative of multiple events of duplication and losses of parts of the cluster.

4.3 Introduction

In filamentous fungi, genes involved in the same secondary metabolite biosynthetic pathway are often located at the same locus in the genome and co-expressed, defining gene clusters (Keller and Hohn 1997). Genomic clustering of genes with related cellular functions (but unrelated sequences) also occurs in other eukaryotes including mammals, nematodes and plants (Hurst et al. 2004; Qi et al. 2004; Shimura et al. 2007). In mammals, it has been shown that clusters of co-expressed genes tend not to be rearranged among species, which indicates that natural selection can act to conserve gene order (Singer et al. 2005; Semon and Duret 2006). Similarly in fungi, natural selection seems to act to conserve gene clusters as exemplified in *Aspergillus* species by the cluster for the biosynthesis of aflatoxin and sterigmatocystin which has been maintained as a cluster, despite many internal rearrangements, for at least 120 million years (Cary et al. 2001; Cary and Ehrlich 2006). The evolutionary mechanisms by which these clusters are created and maintained are unclear, but there is evidence that some instances of clustering result from strong natural selection. For example, the *DAL* cluster involved in nitrogen metabolism in *Saccharomyces cerevisiae* was formed relatively recently by a series of near-simultaneous relocations of genes that were previously scattered around the genome (Wong and Wolfe 2005). Other

mechanisms involved in the formation and maintenance of clusters include selection for co-regulation by chromatin remodeling, epistatic selection for tight linkage between genetically interacting genes, and the "selfish operon" hypothesis of origin by horizontal gene transfer (HGT) (Lawrence 1999; Hurst et al. 2004; Keller et al. 2005; Batada and Hurst 2007; Batada et al. 2007). Indeed, the clustering of the genes from a pathway at a single locus certainly facilitates HGT of genes involved in the same cellular function (Lawrence 1999; Walton 2000) increasing its likelihood.

Despite frequent speculation (reviewed in (Rosewich and Kistler 2000)), and even though some clear examples of HGT of single genes between fungal species (Friesen et al. 2006) or from bacteria to fungi (Wenzl et al. 2005) are known, there are few reports that conclusively demonstrate HGT of a fungal secondary metabolite cluster. The strongest candidate reported so far is the ETP (epipolythiodioxopiperazine) synthase gene cluster, recently analyzed by Patron *et al.* (Patron et al. 2007), but even in this instance alternative evolutionary scenarios can be contemplated (see Discussion). One of the best-known cases of possible HGT of a fungal secondary metabolite cluster concerns the fungal β-lactam (penicillin) antibiotic biosynthetic genes of *Penicillium* species. This proposal was originally made when bacterial and fungal isopenicillin-N-synthetases were found to have unexpectedly highly similar protein sequences (Weigel et al. 1988; Aharonowitz et al. 1992; Liras and Martin 2006). However, subsequent phylogenetic analyses of these proteins failed to provide robust support for their HGT (Smith et al. 1992; Buades and Moya 1996).

The rice blast fungus *Magnaporthe grisea* is one of the richest known fungi in terms of secondary metabolite gene clusters (Dean et al. 2005; Collemare et al. 2008). One of them contains the avirulence gene *ACE1* that encodes a hybrid polyketide synthase - nonribosomal peptide synthetase (PKS-NRPS) likely involved in the biosynthesis of an avirulence signal recognized by rice cultivars carrying the resistance gene *Pi33* (Bohnert et al. 2004). The *ACE1* cluster contains 15 genes that are co-expressed specifically during the appressorium mediated penetration of the fungus into host tissues (J. Collemare *et al.*, submitted). During annotation of the *ACE1* cluster, a similar cluster was identified in the related

animal pathogen *Chaetomium globosum*. We were then interested in identifying possible homologous clusters in other fungi in order to decipher its evolutionary history. In the present study, we combine phylogenetics and comparative genomics to identify orthologs of the *M. grisea ACE1* cluster in other ascomycetes. We define a set of three genes that are shared across all instances of the cluster and hence are probably ancestral to it. This analysis revealed that the cluster in *M. grisea* expanded by internal duplication, and that after this duplication, part of the *ACE1* cluster was likely horizontally transferred from an *M. grisea*-like ancestor into an ancestor of *Aspergillus clavatus*.

4.4 Methods

We set up a local BLAST database of the proteins encoded in 26 completely sequenced fungal genomes (*A. niger*, *A. nidulans*, *A. terreus*, *A. flavus*, *A. oryzae*, *A. clavatus*, *N. fischeri*, *A. fumigatus Af293*, *A. fumigatus CEA10*, *C. immitis*, *C. posadasii*, *P. chrysogenum*, *U. reesii*, *S. sclerotiorum*, *F. graminearum*, *F. oxysporum*, *F. verticillioides*, *M. grisea*, *N. crassa*, *C. globosum*, *H. jecorina* (*T. reesei*), *N. haematococca* (*F. solani*), *P. chrysosporium*, *S. nodorum* (*P. nodorum*), C. *neoformans*, *U. maydis*). To find candidate *ACE1*-like clusters in other fungi, we used a two-step process:

In the first step, each protein encoded by the *M. grisea ACE1* cluster was used as a query in BLASTP searches against this database, and for each query the top 25 hits were retained provided that their E-values were less than 1e-4. Each set of proteins was aligned using ClustalW (Thompson et al. 1994) and poorly aligned regions were removed using Gblocks (Castresana 2000). Sequence alignments can be downloaded from http://wolfe.gen.tcd.ie/nora. Maximum likelihood trees were constructed using PHYML (Guindon and Gascuel 2003) with the JTT amino acid substitution matrix and four categories of substitution rates. Bootstrapping was done using PHYML's default options with 100 replicates per run. To avoid long branch attraction problems we withdrew highly divergent sequences and repeated the alignment and tree reconstruction steps on the new sets. We also verified at each step that the alignment obtained after running Gblocks represented at least

30% of the initial protein sequence. Genes were considered as orthologs of an *M. grisea ACE1* cluster gene if they grouped in a monophyletic clade with a bootstrap support of ≥70%.

Many of the genes identified in this first step were located in gene clusters. For each cluster so identified (defined as the presence of at least two homologs of *M. grisea ACE1* cluster genes adjacent to one another) we then made a second step of analysis, examining any other genes that are physically located within these clusters but which were not picked up at the first step (either because their BLASTP E-values were too weak, or because they were not in the top 25 hits when the database was searched). This process added genes CHG05286.1, CHG05287.1, SNU00307.1 and FVEG_12610 to the analyses.

4.5 Results

4.5.1 Identification of homologous ACE1 clusters in other filamentous fungi

The *ACE1* secondary metabolism gene cluster of *M. grisea* comprises 15 genes: *ACE1* and *SYN2* are PKS-NRPS hybrid genes; *RAP1* and *RAP2* code for enoyl reductases; *CYP1–CYP4* for cytochrome P450 monoxygenases; *ORFZ* for an α/β-hydrolase; *OXR1* and *OXR2* for oxidoreductases; *MFS1* codes for a transporter in the MFS superfamily; *BC2* codes for a binuclear zinc finger transcription factor; *OME1* codes for an O-methyl transferase; and *ORF3* has no homology to known proteins (J. Collemare *et al.*, submitted). To find gene clusters homologous to the *ACE1* cluster in other fungal species, we used an algorithm that searched 26 fungal genomes for loci where at least three likely orthologs of genes from the *ACE1* cluster were linked (see Methods 4.4). This search identified nine similar clusters in seven fungal species from the subphylum Pezizomycotina: three Sordariomycetes (*Chaetomium globosum*, *Fusarium oxysporum* and *F. verticillioides*), one Dothideomycete (*Stagonospora nodorum*) and three Eurotiomycetes (*Aspergillus clavatus*, *Coccidioides immitis* and *Uncinocarpus reesii*) (Figure 4.1).

Two types of clusters related to the *ACE1* cluster were identified: large clusters with eight or more genes are found in *M. grisea*, *C. globosum* and *S. nodorum*, whereas smaller clusters with 3–6 genes are found in the three Eurotiomycetes and in *Fusarium* species (Figure 4.1). *C. globosum* is unusual as its genome contains two large *ACE1*-like clusters which we refer to as clusters #1 and #2. Similarly, the *A. clavatus* genome has two clusters as discussed below. Interestingly, a core set of three genes (homologs of *ACE1*, *RAP1* and *ORF3*; boxed in Figure 4.1 is present in all eight species. The presence of this core suggests that the physical linkage between these three genes is ancient and can be inferred to have existed in the common ancestor of all the genomes considered in Figure 4.1

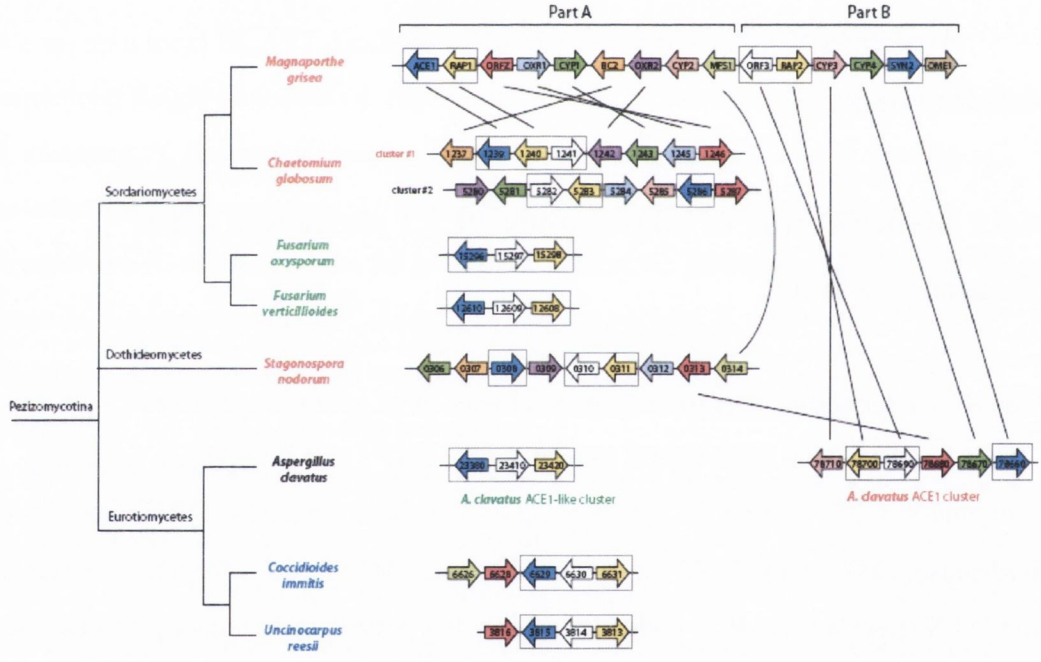


Figure 4.1 *ACE1* and *ACE1*-like gene clusters in filamentous fungi. Colors indicate gene orthology in different species and paralogs in the same species. Horizontal lines indicate genes that are adjacent in the genome, with gene orientations as shown. Genomic regions are not drawn to scale. Parts A and B of the *M. grisea* cluster as identified in the text are marked. The core set of three genes inferred to have been present in the ancestral cluster are boxed. Vertical lines indicate the closest relatives of genes in the *M. grisea* cluster and one of the *A. clavatus* clusters, based on phylogenetic analyses (**Figure 4.3** and **Figure 4.2** The species phylogeny is based on the whole-genome supertree analysis of Fitzpatrick *et al* (Fitzpatrick et al. 2006); in that study the placement of Dothideomycetes relative to Sordariomycetes and Eurotiomycetes varied depending on the method of analysis, so we have shown it as a trichotomy. The analysis of Hane *et al* of the complete *S. nodorum* genome placed Dothideomycetes and Sordariomycetes in a clade with Eurotiomycetes outside (Hane et al. 2007). Species-specific gene nomenclature is shown, except for *M. grisea* (Collemare *et al*, unpublished results). Red, green and blue coloring of species names corresponds to the labelling of individual genes from the clusters in **Figure 4.3** and **Figure 4.2** .

As well as the genes in the eight clusters shown in Figure 4.1 we also identified a small number of single homologs of genes from the *M. grisea ACE1* cluster that are located at dispersed genomic locations in other species.

4.5.2 Phylogenetic analysis of the ACE1 cluster in filamentous fungi

Gene-by-gene phylogenetic analyses were carried out to decipher the evolutionary history of the loci using homologs (even at dispersed locations) of genes from *M. grisea ACE1* cluster (Figure 4.3 and Figure 4.2 The first trend evident from this phylogenetic analysis is that genes from clusters in Eurotiomycetes and *Fusarium* spp. are distant from those of the *M. grisea*, *C. globosum* and *S. nodorum* clusters. Indeed, genes in clusters from these last three species define clades supported by high bootstrap values (> 91%), to the exclusion of genes from Eurotiomycetes and *Fusarium* species (Figure 4.3 a,b,e,f). Interestingly, genes from one of the two clusters in *A. clavatus* are more closely related to genes in the *M. grisea ACE1* cluster than to those in *ACE1*-like clusters from other Eurotiomycetes (see below). In view of the gene contents of the clusters and their phylogenetic relationships, we refer to the large clusters in *M. grisea*, *C. globosum*, *S. nodorum* and the larger of the two clusters in *A. clavatus* as "*ACE1* clusters", and to the smaller clusters in Eurotiomycetes and *Fusarium* spp. as "*ACE1*-like clusters". These two types of cluster have probably had a long history of independent evolution, although they certainly share a common ancestor.

We then focused on the origins of the duplicated genes in the *M. grisea* cluster. Phylogenetic trees show clearly that in *M. grisea RAP2* is a paralog of *RAP1*,

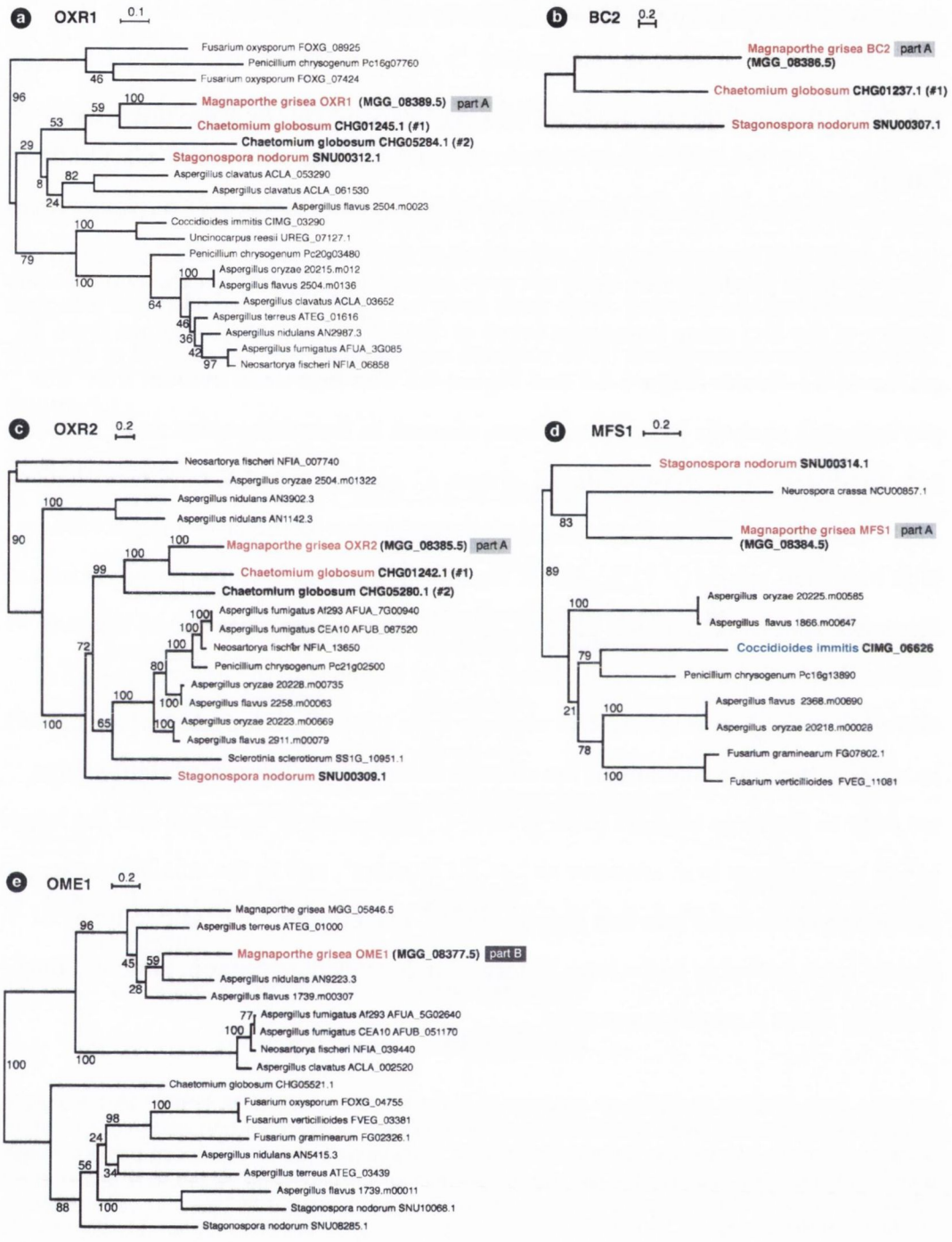


Figure 4.2 Maximum-likelihood trees for *ACE1* cluster genes and their homologs. Trees were constructed as for **Figure 4.3** The values of the shape parameter (alpha) for the gamma distribution were estimated from the data as 1.516, 1.482, 1.568, 2.826 and 2.483 for panels *a-e*, respectively. The proportions of invariant sites are 0.013, 0.028, 0.004, 0.017 and 0.019, respectively.

CYP3 is a paralog of *CYP2*, *CYP4* is a paralog of *CYP1*, and *SYN2* is a paralog of

ACE1 (Figure 4.3 a-d). Notably, in each of these pairs, one gene is located on the left-hand side of the *M. grisea* cluster and the other is on the right-hand side. Thus the *M. grisea* cluster appears to have undergone partial tandem duplication at some stage during its evolution, although the gene order is not conserved between the two parts. The presence of two *ACE1* clusters in *C. globosum* is suggestive of a second block-duplication event in this species. However, for most genes present in both *C. globosum ACE1* clusters, the copy from cluster #1 forms a clade with their *M. grisea* homologs. This close phylogenetic relationship is observed for *ACE1*, *RAP1*, *ORFZ*, *OXR1*, *CYP1,* and *OXR2*. The only exception to this pattern is *M. grisea ORF3* which is marginally closer to the *C. globosum* cluster #2 gene, but with low bootstrap support (Figure 4.3 and Figure 4.2 This observation suggests that the duplication that gave rise to the current *C. globosum* clusters #1 and #2 occurred in a common ancestor of *C. globosum* and *M. grisea*, and that the corresponding cluster #2 in *M. grisea* was lost.

On the basis of this analysis, we divided the *M. grisea* cluster into two parts, A and B, so that each of the duplicated genes in *M. grisea* has one copy in Part A and one in Part B (Figure 4.1 Part A in *M. grisea* consists of nine genes, all of which have orthologs in one or both of the clusters in its closest relative *C. globosum*. The clusters in other species consist of homologs of genes from *M. grisea* Part A, plus one gene from Part B (*ORF3*; see Discussion 4.6). The order of the Part A genes is not conserved among *M. grisea*, *C. globosum* and *S. nodorum*.

Surprisingly, this phylogenetic analysis shows that five of the six genes from Part B of the *M. grisea ACE1* cluster group with genes from the larger of the two clusters in *A. clavatus*, rather than with the genes in the more closely related (Sordariomycete) species *C. globosum*, or with their Part A paralogs in *M. grisea*. Bootstrap values for grouping the *M. grisea* Part B genes *SYN2*, *RAP2*, *CYP4*, *CYP3* and *ORF3* with their *A. clavatus* homologs are 98-100% (Figure 4.3 a-e). The only gene from Part B of the *M. grisea* cluster that does not group with *A. clavatus* is *OME1* (Figure 4.2 e), but this is also the only gene whose detected homolog in *A. clavatus* (*ACLA_002520*) is not physically clustered with the others, which calls its orthology into question. The consistency of this

phylogenetic result for Part B genes, and its disagreement with the expected species relationships, are indicative of HGT between *A. clavatus* and Part B of the *M. grisea* cluster. In contrast seven of the nine genes from Part A of the *M. grisea*

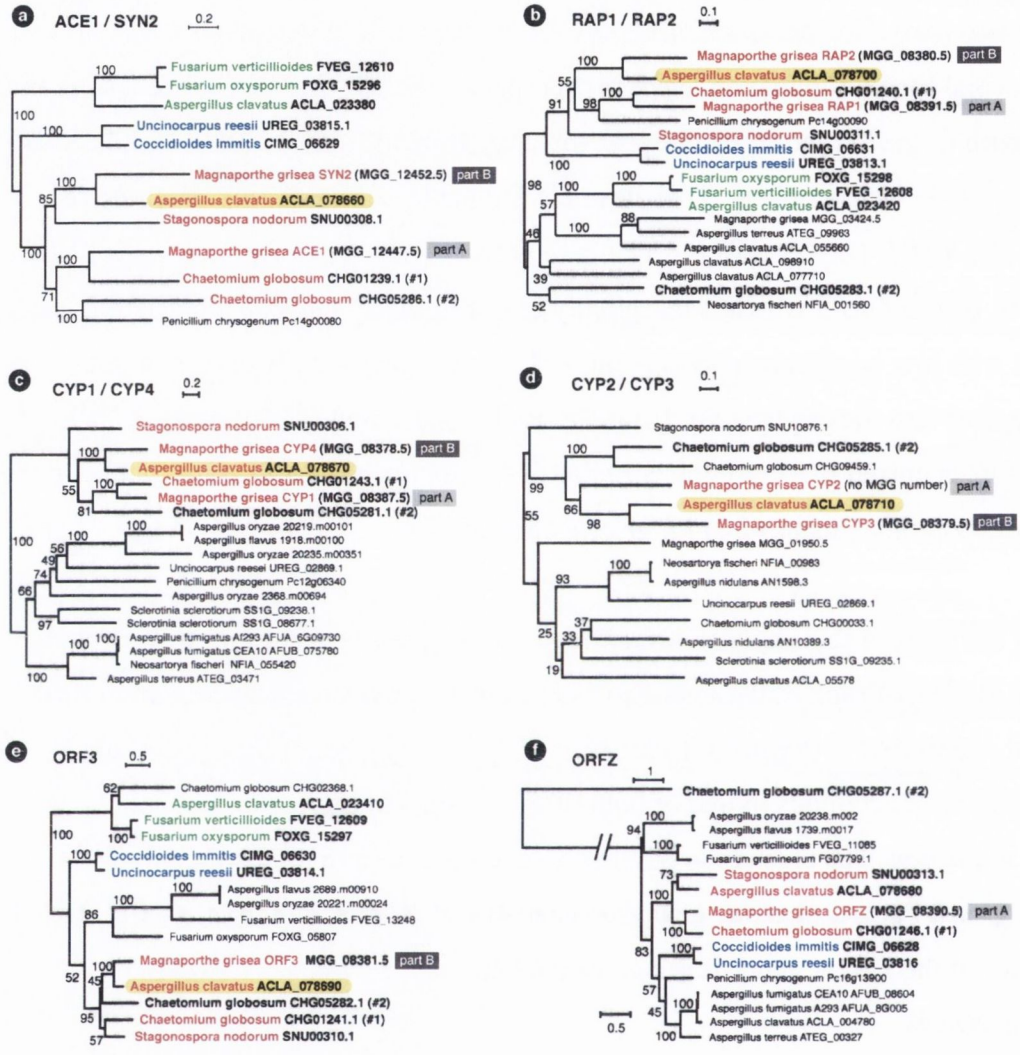


Figure 4.3 Maximum likelihood trees for *ACE1* cluster genes and their homologs. (a) *ACE1* and *SYN2*; (b) *RAP1* and *RAP2*; (c) *CYP1* and *CYP4*; (d) *CYP2* and *CYP3*; (e) *ORF3*; (f) *ORFZ*. In each tree, genes that appear in Figure 1 are named in color or bold black. Yellow highlighting shows the five genes in the *A. clavatus ACE1* cluster whose closest relatives are genes from part B of the *M. grisea* cluster. Bootstrap percentages are shown for all nodes. Trees were constructed from amino acid sequences as described in Methods using PHYML after alignment with ClustalW and Gblocks filtering. Trees for the other five genes in the *ACE1* cluster are shown in Additional data file 1. The values of the shape parameter (α) for the gamma distribution were estimated from the data as 1.329, 1.441, 2.476, 2.615, 2.536 and 0.961 for panels a-f, respectively. The proportions of invariant sites are 0.028, 0.035, 0.030, 0.068, 0.000 and 0.000, respectively. The *M. grisea SYN2* gene corresponds to parts of the automatically-annotated gene models MGG_12452.5 and MGG_12451.5.

cluster, including *ACE1* itself, lie at the expected phylogenetic position forming a clade with *C. globosum* (Figure 4.3 and Figure 4.2 ; the two exceptions are *CYP2* which is discordant but has a low bootstrap value of 66%, and *MFS1* which cannot be analyzed because there is no homolog in the *C. globosum* clusters).

For the four panels in Figure 4.3 that include sequences from other Eurotiomycetes (*C. immitis* and *U. reesii*) as well as *A. clavatus*, we used the Likelihood Ratio Test (LRT) to test whether the topologies shown (Figure 4.3 a,b,e,f) have significantly higher likelihoods than alternative trees where the Eurotiomycetes were constrained to form a monophyletic group. In all four cases the topology shown in Figure 4.3 is significantly more likely than the tree expected if genes were inherited vertically ($P < 0.001$ for each).

4.5.3 Identifying the direction of gene transfer

To determine whether Part B of the cluster was transferred from an *M. grisea*-like donor to an ancestor of *A. clavatus*, or vice versa, we examined phylogenetic trees constructed from those genes that have orthologs both in species that are close relatives of *M. grisea* and in species that are closer to *A. clavatus*. We would predict that if an ancestor of *A. clavatus* was the recipient of HGT, then the genes in its *ACE1* cluster would not show the expected close relationship to other Eurotiomycete species such as *C. immitis* and *U. reesii* (Figure 4.1 and would instead form a clade with the donor lineage (represented by *M. grisea*). Conversely, if the direction of transfer was from an *A. clavatus*-like donor into the *M. grisea* lineage, we would expect the *M. grisea* Part B genes not to form a monophyletic clade with the other Sordariomycete species *C. globosum*, and instead to group with *A clavatus*.

In the phylogenetic tree of *ORF3* sequences, the shared *A. clavatus–M. grisea* branch lies within a clade that contains homologs from the two clusters in *C. globosum*, as well as the Dothideomycete *S. nodorum* (Figure 4.3 e). The *ORF3* orthologs from *C. immitis* and *U. reesii* clearly lie outside this clade with 95% bootstrap support. Similarly, the phylogenetic tree of *RAP1* and *RAP2* orthologs (Figure 4.3 b) shows that the shared branch containing the *A. clavatus* gene and the Part B *M. grisea* gene (*RAP2*) lies within a larger clade that includes the *C. globosum* and *M. grisea* Part A (*RAP1*) orthologs. The homologs from *C. immitis* and *U. reesii* lie outside (91% bootstrap support). Likewise, the

phylogenetic tree of the *ACE1–SYN2* pair (Figure 4.3 a) places the *A. clavatus* sequence within a Sordariomycete/Dothideomycete clade, distant from the other Eurotiomycetes (*C. immitis* and *U. reesii*). These topologies all indicate that an ancestor of *M. grisea* was the donor of the transferred Part B genes, and an ancestor of *A. clavatus* was the recipient.

ORFZ is the only gene in the *A. clavatus ACE1* cluster that does not have a homolog in Part B of the *M. grisea* cluster. The origin of this gene in *A. clavatus* is not clear. Phylogenetic analysis (Figure 4.3 f) indicates that *A. clavatus ORFZ* does not group with the *C. immitis* and *U. reesii* genes, and this conclusion is supported by the Likelihood Ratio Test. This result suggests a foreign origin for *A. clavatus ORFZ*, but the absence of a homolog in *M. grisea* Part B makes it impossible to test whether this gene has a similar origin to its five neighboring genes in *A. clavatus*.

We conclude that there is phylogenetic support for the hypothesis that at least five of the six genes in the *ACE1* cluster of *A. clavatus* originated by HGT, and that the most probable single donor is a Sordariomycete ancestor related to *M. grisea*.

4.6 Discussion

4.6.1 The ACE1 cluster is specific to few fungal species

A complete *ACE1* cluster is present in only four of the 23 sequenced Pezizomycotina genomes (*M. grisea*, *C. globosum*, *S. nodorum* and *A. clavatus)*. Such a sporadic distribution could be the result of either independent HGTs or frequent losses of the whole cluster in different lineages (Figure 4.4). We favor the latter explanation because – with the exception of *A. clavatus* – our phylogenetic trees of genes from the cluster have topologies that are in broad agreement with the expected species phylogeny (Fitzpatrick et al. 2006). We suggest that an *ACE1*-like cluster consisting of at least three genes (homologous to *ACE1*, *RAP1* and *ORF3*) existed in the common ancestor of Pezizomycotina, but this cluster has been lost in many lineages subsequently. The scheme in Figure

4.4 identifies four independent lineages (shown by dashed lines) in which all copies of the cluster have been lost. We cannot tell, with current data, whether genes such as *OXR1* that are present in the *ACE1* clusters of Sordariomycetes and Dothideomycetes but not in the *ACE1*-like clusters of Eurotiomycetes correspond to lineage-specific additions or losses.

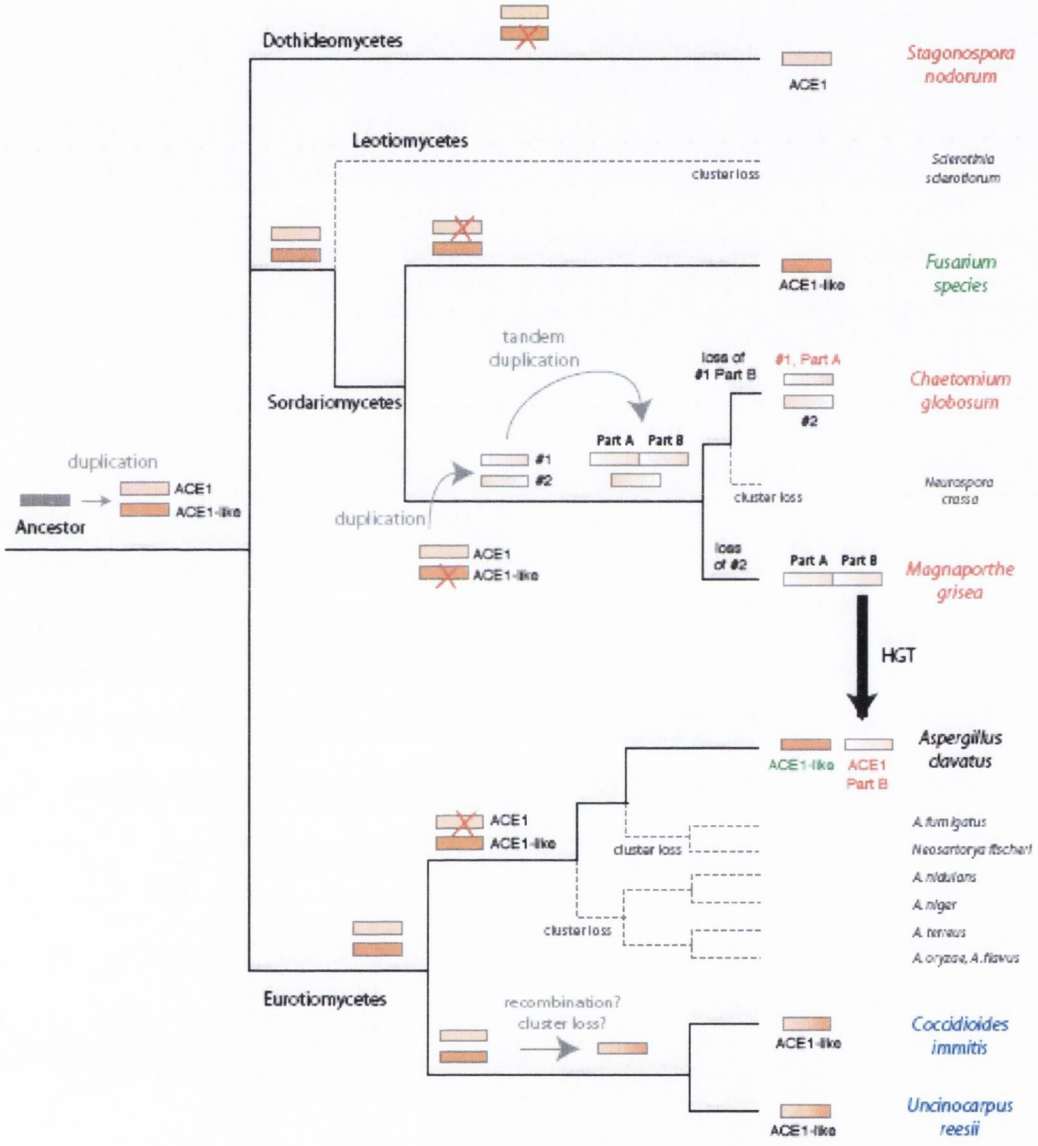


Figure 4.4 Inferred history of *ACE1* and *ACE1*-like clusters in filamentous fungi. The gray rectangle corresponds to the ancient core cluster of three genes (*ACE1*, *RAP1*, *ORF3*) that is common to all *ACE1* clusters (pink) and *ACE1*-like clusters (orange). The black arrow denotes the inferred HGT of part B of the cluster from a donor related to *M. grisea* to the *A. clavatus* recipient. Dashed branches and smaller fonts indicate euascomycetes that were included in our analysis but lack the clusters entirely. Phylogenetic relationships are based on (Fitzpatrick et al. 2006) and Chapter 7, for the topology within the genus *Aspergillus*. The tree is not drawn to scale.

Any tree showing apparent HGT of a gene can also be explained by an alternative scenario of gene duplications and losses. However, the situation reported here is

rather different to typical cases of possible HGT of individual genes, because it involves multiple genes that are arranged as a large tandem duplication (in *M. grisea*). The fact that the *A. clavatus ACE1* cluster forms a clade with the *M. grisea* Part B genes (to the exclusion of the Part A genes) means that the only alternative scenario to HGT is one where the Part A/Part B tandem duplication occurred right at the base of the tree in Figure 4.4 This scenario would then necessitate at least four events of precise loss of exactly one part of the tandemly duplicated set of genes: Part B in *C. globosum*, Part B in the ancestor of *C. immitis* and *U. reesii*, Part B in *S. nodorum*, and Part A in *A. clavatus.* Because of the precise nature of the deletion required (and choice of gene copy to delete), we do not regard this scenario as likely.

The discontinuous distribution of the *ACE1* cluster among fungal species suggests that evolutionary constraints act to maintain this cluster only in few species. As *M. grisea*, *S. nodorum* and *C. globosum* are plant or animal pathogens, it is tempting to speculate that the *ACE1* cluster is involved in the infection process of these three species. The metabolite produced by this biosynthetic pathway may be an important pathogenicity factor, but such a role remains to be determined. *A. clavatus* is different as it is not pathogenic. The presence of the *ACE1* cluster in *A. clavatus* may arise from selection involving an unknown biological role of this metabolite in this fungus. Identifying the molecules made by these different clusters will be necessary to understand the role of the *ACE1* cluster in fungal biology and could give clues about evolution of the ancestral biosynthetic pathway-controlled by this cluster.

4.6.2 ACE1 cluster evolution in Sordariomycetes involved several duplication events

The *ACE1* cluster has a complex history with multiple events of large-scale duplication and multiple losses. The scenario we infer is summarized in Figure 4.4 An ancient duplication produced the large *ACE1* and smaller *ACE1*-like clusters. A second duplication event in an ancestral Sordariomycete gave rise to the two clusters (#1 and #2) presently seen in *C. globosum.* This event occurred prior to

the speciation between *C. globosum* and *M. grisea*, but *M. grisea* later lost its counterpart of cluster #2. Independently, cluster #1 underwent a tandem duplication event, generating Parts A and B. This tandem duplication survived in *M. grisea*, but in *C. globosum* the addition (Part B of cluster #1) was lost again. It might seem simpler to suggest that the Part A/B tandem duplication was an event that occurred specifically in *M. grisea* after it diverged from *C. globosum*, but we know that this is incorrect because the Part B genes from *M. grisea* form outgroups to a clade consisting of *C. globosum* and *M. grisea* Part A genes. We can also be sure that the surviving duplications seen in *M. grisea* and *C. globosum* were separate events because of the topology of the phylogenetic trees: if the surviving genes were descended from the same duplication event we would expect that in the *ACE1-SYN2* tree, for example, *M. grisea ACE1* and *SYN2* should each form a separate monophyletic group with one of the *C. globosum* genes, but that is not seen (Figure 4.3 a). Instead we interpret the trees as indicative of two duplications of the whole cluster in a Sordariomycete ancestor of *M. grisea* and *C. globosum*, the first of which was non-tandem and the second of which was tandem. After this tandem duplication, the *M. grisea* lineage lost its ortholog of *C. globosum*'s cluster #2, and the *C. globosum* lineage lost its ortholog of *M. grisea*'s Part B (Figure 4.4 This pattern of frequent loss is consistent with the cluster's sporadic distribution in fungi.

ORF3 is unusual as it is inferred to have been present in the ancestor of all *ACE1* and *ACE1*-like clusters, but in *M. grisea* it is not duplicated and it shows phylogenetic affinity to *A. clavatus* rather than to *C. globosum* or *S. nodorum* (Figure 4.2 e). These properties suggest that a homolog of *ORF3* was lost from Part A of the *M. grisea* cluster, after the tandem duplication occurred. Furthermore, we speculate that *ORF3*'s location on the boundary between Parts A and B may indicate that the tandem duplication event visible in *M. grisea* involved a recombination between two copies of this gene.

Gene order and orientation is quite poorly conserved among the *ACE1* clusters, as is typical of many secondary metabolism gene clusters (Cary et al. 2001; Gardiner et al. 2004; Cary and Ehrlich 2006). This makes it all the more striking that the duplicated *M. grisea* genes each have one copy in the Part A and one copy in Part

B. Because the tandem duplication that is evident in the *M. grisea* genome is not particularly recent (it predates the *M. grisea*/*C. globosum* speciation), we suggest that some form of selection has acted on gene order in the cluster, preventing intermixing of the two parts. In this context it is notable that recombination seems to be inhibited in the *M. grisea ACE1* cluster, because it displays a low frequency of targeted gene replacement, even in a *KU80* null mutant background where homologous recombination rates are increased ((Villalba et al. 2008); J. Collemare *et al.*, submitted). The way that Part A and Part B genes of the *ACE1* cluster are distributed among species may indicate that they are involved in the biosynthesis of different molecules. Alternatively, Parts A and B of the *ACE1* cluster may be each involved in the biosynthesis of independent polyketide precursors that are fused into a final complex molecule as observed for lovastatin (Kennedy et al. 1999; Hutchinson et al. 2000; Collemare et al. 2008). The fact that all 15 genes in the *M. grisea ACE1* cluster are co-expressed at a very specific stage of the infection process (J. Collemare *et al.*, submitted) favors the hypothesis that both Part A and Part B genes are involved in same biosynthetic pathway. However, gene knockout experiments have shown that two Part B genes (*RAP2* and *SYN2*) are not essential for the avirulence function supported up to now only by the Part A gene *ACE1* (J. Collemare *et al.*, submitted). These latter results suggest that Part A and Part B genes could be involved in the biosynthesis of two different molecules, with only one (*ACE1*, Part A pathway) being recognized by resistant rice cultivars. However, these two hypotheses are both plausible, and await the biochemical characterization of the Ace1 metabolite.

4.6.3 HGT of a fungal secondary metabolism gene cluster

Although the genomics era has uncovered evidence for widespread horizontal gene transfer among prokaryotes (Doolittle 1999; Beiko et al. 2005), and from prokaryotes to eukaryotes (Kondo et al. 2002; Wenzl et al. 2005; Woolfit et al. 2007) or *vice versa* (Guljamow et al. 2007; Rogers et al. 2007), relatively few instances of horizontal gene transfer have been documented from one eukaryote to another (Andersson et al. 2005; Richards et al. 2006; Andersson et al. 2007). Among fungi, the best documented is the transfer of a virulence gene from

S. nodorum to *Pyrenophora tritici-repens*, which occurred only about 70 years ago (Friesen et al. 2006). In that case, the transferred DNA fragment was about 11 kb in size but contained only one gene. In this study we showed that Part B of the *ACE1* cluster (30 kb in size, containing 5-6 genes) was likely horizontally transferred from a close ancestor of *M. grisea* (a Sordariomycete) into an ancestor of *A. clavatus* (a Eurotiomycete). The mechanism by which HGT might have occurred remains a matter of speculation, but could perhaps have involved hyphal fusion between species, or endocytosis. Our inference of HGT is valid only if the Sordariomycete and Eurotiomycete clades are monophyletic as shown in Figure 4.1 but their monophyly is supported by several molecular and systematics analyses (Blackwell et al. 2006; Fitzpatrick et al. 2006; James et al. 2006; Kuramae et al. 2006; Robbertse et al. 2006; Hane et al. 2007).

To our knowledge, our study and the recent work of Patron *et al.* (Patron et al. 2007) are the first reported instances of HGT of groups of linked genes involved in the same pathway between eukaryotic species. In both cases these secondary metabolite clusters show a punctate (sporadic) distribution among other species, with an ancestral cluster apparently having been lost by more species than the number that retain it. This pattern of frequent losses of genes and their occasional reacquisition by HGT resembles the pattern of evolution of "dispensable pathway" genes in ascomycete yeasts (Hall and Dietrich 2007). Hall and Dietrich (Hall and Dietrich 2007) noted that genes whose products function in dispensable pathways are one of the few categories of genes in *S. cerevisiae* that are physically organized into gene clusters. They found that the pathway for biotin synthesis was lost in a yeast ancestor and then regained in the *S. cerevisiae* lineage by a combination of HGT from bacteria and gene duplication with neofunctionalization. One possible explanation for this strange pattern of evolution could be that an intermediate in the pathway is toxic (Hall and Dietrich 2007), although there is no direct experimental evidence of this. If a pathway can confer a selective advantage in some circumstances but also involves the production of a toxic intermediate, there can be strong selection in favor of the pathway in some conditions and strong selection against it in others. The consequences of such a situation could include the formation of physical gene clusters (to reduce the chances of coding for only part of the pathway, or for

strong repression of transcription mediated by chromatin remodeling), and occasional selection for re-gain of function by HGT. Further exploration of this hypothesis will require the discovery of more examples of similar sets of genes, and detailed characterization of the biochemical pathways involved.

Chapter 5

Elusive Origins of the Extra Genes in *Aspergillus oryzae*

This chapter is based on a manuscript submitted to *PLOS ONE* (authors **N. Khaldi** and K.H. Wolfe). We are currently revising the manuscript in response to the reviewers' comments.

5.1 Abstract

The genome sequence of *Aspergillus oryzae* revealed unexpectedly that this species has approximately 20% more genes than its congeneric species *A. nidulans* and *A. fumigatus*. Where did these extra genes come from? Here, we evaluate several possible causes of the elevated gene number. Many gene families are expanded in *A. oryzae* relative to *A. nidulans* and *A. fumigatus*, but we find no evidence of ancient whole-genome duplication or other segmental duplications, either in *A. oryzae* or in the common ancestor of the genus *Aspergillus*. We show that the presence of divergent pairs of paralogs is a feature peculiar to *A. oryzae* and is not shared with *A. nidulans* or *A. fumigatus.* In phylogenetic trees that include paralog pairs from *A. oryzae*, we frequently find that one of the genes in a pair from *A. oryzae* has the expected orthologous relationship with *A. nidulans*, *A. fumigatus* and other species in the subphylum Eurotiomycetes, whereas the other *A. oryzae* gene falls outside this clade but still within the Ascomycota. We identified 456 such gene pairs in *A. oryzae*. Further phylogenetic analysis did not however indicate a single consistent evolutionary origin for the divergent members of these pairs. Approximately one-third of them showed phylogenies that are suggestive of horizontal gene transfer (HGT) from Sordariomycete

species, and these genes are closer together in the *A. oryzae* genome than expected by chance, but no unique Sordariomycete donor species was identifiable. The postulated HGTs from Sordariomycetes still leave the majority of extra *A. oryzae* genes unaccounted for, and one possible explanation for our observations is that *A. oryzae* might have been the recipient of many separate HGT events from diverse donors.

5.2 Introduction

Aspergillus oryzae is a filamentous ascomycete used in traditional Japanese soy sauce production, and has important industrial applications as a producer of hydrolytic enzymes in solid-state fermentations (Hamada et al. 1992; Kobayashi et al. 2007). When the *A. oryzae* genome was sequenced it was found to be significantly larger and to contain more genes than the genomes of other species in the genus *Aspergillus* (Galagan et al. 2005; Machida et al. 2005; Nierman et al. 2005). At 37 Mb, the *A. oryzae* genome assembly is 23% larger than the *A. nidulans* assembly and 32% larger than the *A. fumigatus* assembly. It is predicted to code for 12,074 proteins of >100 amino acids, which is 1,412 more proteins than *A. nidulans* and 2,444 more than *A. fumigatus* (Galagan et al. 2005). Although some of the differences in the predicted gene count between species may be due to differences in the bioinformatics methods used by the three sequencing laboratories to annotate the genomes, the larger DNA content of *A. oryzae* is incontrovertible. Moreover, even if one excludes singleton genes (those with no homologs in completed fungal genome databases), which are the ones most likely to be annotation artifacts, *A. oryzae* still contains 16-26% more genes than the other species: the numbers of non-singletons are 12,044, 10,425 and 9,574 for *A. oryzae*, *A. nidulans* and *A. fumigatus* respectively. There is also indirect evidence that most of the predicted genes in *A. oryzae* are functional: sequencing the genome of *A. flavus*, which is a very close relative of *A. oryzae* and has a similarly large genome, shows that 8,953 genes (85% of all orthologs between *A. oryzae and A. flavus*) have a conservative pattern of nucleotide substitution ($K_A < K_S$) consistent with purifying selection to retain protein-coding capacity (**Error! Reference source not found.**, unpublished results).

Comparisons of gene order revealed the presence of syntenic blocks common to *A. oryzae*, *A. nidulans* and *A. fumigatus*, as well as genomic blocks specific to *A. oryzae* that lack synteny with the other two species (Machida et al. 2005). These non-syntenic blocks share the characteristics of, first, containing genes that seem to exist only in *A. oryzae* (that is, they lack orthologs in *A. nidulans* and *A. fumigatus*) and, second, appearing in a mosaic manner throughout the genome of *A. oryzae*. Machida *et al.* (Machida et al. 2005) reported that, surprisingly, phylogenetic analysis of some genes from these non-syntenic regions indicated that they are distantly related paralogs of other genes that have a syntenic and apparently orthologous relationships among *A. oryzae*, *A. nidulans* and *A. fumigatus*. Thus, the extra genes in *A. oryzae* did not seem to be the result of recent gene duplications that happened specifically in *A. oryzae* after it diverged from the other two species; rather, they seemed to be the products of much older divergence event(s) (Machida et al. 2005).

In this study we refer to this distinctive phylogenetic tree topology, first commented upon by Machida *et al.* (Machida et al. 2005), as the S topology (Figure 5.1 We show that the discovery of these trees is not just a subjective observation: there is a genuine statistical excess of trees with the S topology in *A. oryzae* as compared to other *Aspergillus* species. We then explore several possible explanations for these trees. First, we test whether the pattern could have been caused by an ancient whole-genome duplication (WGD) in the common ancestor of the three *Aspergillus* species, similar to the WGD that occurred in an ancestor of *Saccharomyces cerevisiae* (Wolfe and Shields 1997; Kellis et al. 2004), followed by a lower rate of gene loss in *A. oryzae* than in

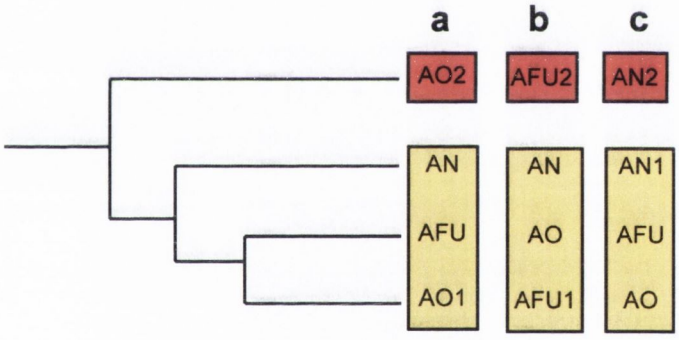


Figure 5.1 Definition of Topology S loci in (a) *A. oryzae*, (b) *A. fumigatus*, and (c) *A. nidulans*. Yellow boxes indicate the genes that are orthologous in the three species. The red box contains the extra paralog (labeled *AO2*, *AFU2* or *AN2*). Loci having Topology S were identified in each species by an automated process without the use of an outgroup, but the root of the tree is likely to lie on the upper (longest) branch.

A. nidulans and *A. fumigatus*. We find no evidence for WGD or similar events such as whole-chromosome duplication (aneuploidy). Second, we consider the possibility that the Topology S trees might be artifacts. If the paralogous pairs in *A. oryzae* were in fact the products of recent (species-specific) gene duplications, but one gene copy subsequently underwent rapid sequence divergence, the ensuing asymmetry of evolutionary rates could cause phylogenetic methods to infer an incorrect tree due to the phenomenon of long-branch attraction between the accelerated branch and the outgroup (Felsenstein 1978; Fares et al. 2006). Third, we consider the possibility that the extra genes were added to the *A. oryzae* genome by horizontal gene transfer (HGT).

5.3 Methods

5.3.1 Data

Genome sequences and annotations of *A. nidulans*, *F. graminearum* and *U. maydis* were downloaded from the Broad Institute (http://www.broad.mit.edu/annotation/fgi/); *A. fumigatus* from TIGR (http://www.tigr.org/tdb/e2k1/afu1/); and *A. oryzae* from the DDBJ database (accession numbers AP007150-AP007177). BLAST searches against the *A. flavus* genome were performed on http://gaplabg5.cfr.ncsu.edu/blast/blast.html.

5.3.2 Identification of Sets S_{AO}, S_{AN} and S_{AFU}

We accepted as triplet sets of orthologous genes in *A. oryzae*, *A. nidulans* and *A. fumigatus*, those genes that were best mutual BLASTP hits. We then tested whether a paralog of the *A. oryzae* gene was present. To identify pairs of paralogs in *A. oryzae* that satisfy topology S, we required that (*i*) the *A. nidulans* and *A. fumigatus* genes in the triplet had no paralogs, and (*ii*) the amino acid sequence divergences between the three orthologs in the triplet were all shorter than the distance between each of the orthologs and the *A. oryzae* paralog. If these conditions were met, the genes were retained for further analysis and the *A. oryzae* orthologous and paralogous copies were termed *AO1* and *AO2*

respectively. An automated search using these criteria resulted in the Set S_{AO} of 456 loci that are duplicated only in *A. oryzae* and have topology S. Analogous methods were used to identify Sets S_{AN} and S_{AFU}.

5.3.3 Testing for WGD

To evaluate the hypothesis that the pairs of genes in Set S_{AO} could be the products of a whole-genome duplication, we tested whether these pairs tend to be located on sister genomic regions. If a WGD occurred and produced two pairs of duplicates, A' and A", B' and B", we would expect that if A' and B' are physically close together in the genome, their paralogs A" and B" should also be close together somewhere else in the genome. We examined whether the Set S_{AO} paralogs in *A. oryzae* genome tend to be arranged in this pattern, relative to random expectations. In this analysis we did not take into account the distinction between the *AO1* and *AO2* members of the *A. oryzae* pairs (which have and do not have, respectively, orthologs in *A. nidulans* and *A. fumigatus*).

We used a sliding window approach to find regions of relatively high paralog density in *A. oryzae*. We considered windows of size *j* genes (*j* = 200 genes in the examples shown in Figure 5.4 and counted how many genes in Set S_{AO} they contained. We then retained only the denser windows for further analysis: those with some number *i* genes in Set S_{AO} (*i* = 6 in Figure 5.4 C). For each window with a particular value of *i*, we then tried to measure the physical distance (in units of genes) spanned by their paralogs elsewhere in the *A. oryzae* genome. In many cases the paralogs were not all on the same chromosome. For those windows whose paralogs were all on the same chromosome, we computed the total chromosomal distance *d* occupied by the paralogous windows. We then evaluated the statistical significance of the observed value of *d* by comparing it to the empirical distribution of values obtained in 10,000 computer simulations where the locations of the paralogous genes were randomized. We carried out a parallel analysis on *S. cerevisiae* as a positive control, using the 551 gene pairs from Byrne and Wolfe (Byrne and Wolfe 2005) in place of Set S_{AO}. In a second analysis, we also included data from Set S_{AO} genes whose paralogs were located on different *A. oryzae* chromosomes, by adding a fixed penalty to the *d* value in

these cases (Figure 5.4 D). The penalty was set at 3000, which is approximately the average number of genes on an *A. oryzae* chromosome. Experimenting with different values of *j* and *i* (data not shown) did not reveal any statistically significant evidence of WGD for *A. oryzae*, whereas the results for *S. cerevisiae* were consistently significant.

5.3.4 Testing for HGT

We used the basidiomycete *U. maydis* as an outgroup for all phylogenetic analyses. Among the 456 loci in Set S_{AO}, we retained a subset (Q_{AO}) of 122 loci for which the same *U. maydis* protein was identified as the reciprocal best BLASTP hit of the S_{AO} set members from all three *Aspergillus* species. We then used BLASTP to identify additional fungal homologs of the Set Q_{AO} genes pairs in the NCBI nonredundant protein sequence database. Proteins were aligned using ClustalW (Thompson et al. 1994), and Gblocks (Castresana 2000) was used to remove poorly aligned positions and regions. Maximum likelihood trees were constructed using PHYML (Guindon and Gascuel 2003) with the JTT amino acid substitution matrix and four categories of substitution rates. Bootstrapping was done using PHYML's default options with 100 replicates per run. The trees for each gene were inspected manually to detect common themes and classified into the three possible mutually exclusive topologies (Figure 5.5 Analogous methods were used to make and study Sets Q_{AN} and Q_{AFU}.

5.4 Results

5.4.1 An excess of topology S trees

Machida *et al.* (Machida et al. 2005) noted that many gene families are expanded in *A. oryzae* and that phylogenetic trees of these families often show the topology that we refer to as Topology S. This topology is remarkable because it implies that a pair of paralogous genes (*AO1* and *AO2*; Figure 5.1 a) that are present in *A. oryzae* originated by a gene duplication that occurred before the speciation events that separated *A. oryzae*, *A. nidulans* and *A. fumigatus*. If that interpretation

is correct, then orthologs of the *AO2* gene must subsequently have been deleted from the *A. nidulans* and *A. fumigatus* genomes, and the shape of the species phylogeny necessitates that two independent deletions of *AO2* must have occurred in *A. nidulans* and *A. fumigatus*.

We quantified this phenomenon by searching systematically for loci that form a paralogous pair in *A. oryzae* but are single-copy in *A. nidulans* and *A. fumigatus*, and where one of the *A. oryzae* copies is divergent from the other three sequences as expected under Topology S (see Methods 5.3). We found 456 such pairs, which accounts for 19% of all pairs of paralogous genes in *A. oryzae*. We refer to these loci as Set S_{AO}. For comparison, we likewise defined two analogous sets of loci that contain extra copies in other species. Set S_{AFU} consists of genes that are double-copy only in *A. fumigatus* and have the equivalent of Topology S for that species (Figure 5.1 b). This set contains 202 pairs, which is 11% of all paralog pairs in *A. fumigatus*. Similarly, Set S_{AN} consists of 219 pairs (12% of all paralog pairs) that are duplicated only in *A. nidulans* and have the topology shown in Figure 1c.

The number of loci in Set S_{AO} is significantly greater than the numbers in Set S_{AN} ($P = 2 \times 10^{-10}$ by Fisher test) and Set S_{AFU} ($P = 7 \times 10^{-12}$), whereas there is no significant difference between Sets S_{AN} and S_{AFU} ($P = 0.43$). This result confirms Machida *et al.*'s report that there is an excess of divergent paralogous gene copies in the *A. oryzae* genome.

5.4.2 Topology S trees are not artifacts of rate acceleration after gene duplication

Gene duplication is often followed by a period of accelerated sequence evolution in one or both gene copies (Lynch and Conery 2000; Van de Peer et al. 2001; Conant and Wagner 2003). If the rates are sufficiently unequal, the resulting phylogenetic trees can suffer from long branch attraction (Felsenstein 1978; Van de Peer et al. 2001; Conant and Wagner 2003), an artifact of tree reconstruction methods that causes the longer branches to clump together in the tree regardless of their true phylogenetic relationship. Where one member of a duplicated gene pair

has accelerated, attraction between this gene's branch and the outgroup used to root the tree (which is usually also a long branch) can result in an incorrect topology that makes the gene duplication appear older than it actually was. In the case of *Saccharomyces cerevisiae*, this artifact affected more than half of all the duplicate gene pairs that were formed by its WGD (Langkjaer et al. 2003; Fares et al. 2006).

To test whether the *A. oryzae* Topology S trees are the result of a similar artifact, we compared the levels of synonymous nucleotide divergence (K_S) between each of the *A. oryzae* genes (*AO1* and *AO2*) and their single homolog (*AN*) in *A. nidulans*. Rate acceleration after gene duplication is expected to affect synonymous divergence to a much lesser extent than nonsynonymous divergence, if at all. If the duplication that produced *AO1* and *AO2* was a species-specific event in *A. oryzae* then the distances |*AO1-AN*| and |*AO2-AN*| should be equal, but if *AO2* branched off before the *A. oryzae-A. nidulans* divergence then |*AO1-AN*| should be less than |*AO2-AN*| (Note that *AO2* is defined as the *A. oryzae* gene that lacks an apparent *A. nidulans* ortholog; Figure 5.1 a). We find that the synonymous divergence is indeed higher for the *AO2* genes: the median K_S values are 3.35 for |*AO2-AN*| and 2.87 for |*AO1-AN*| ($P < 10^{-4}$ by Wilcoxon test; Figure 2). This result indicates that the excess of genes in Set S_{AO} is not an artifact of long branch attraction.

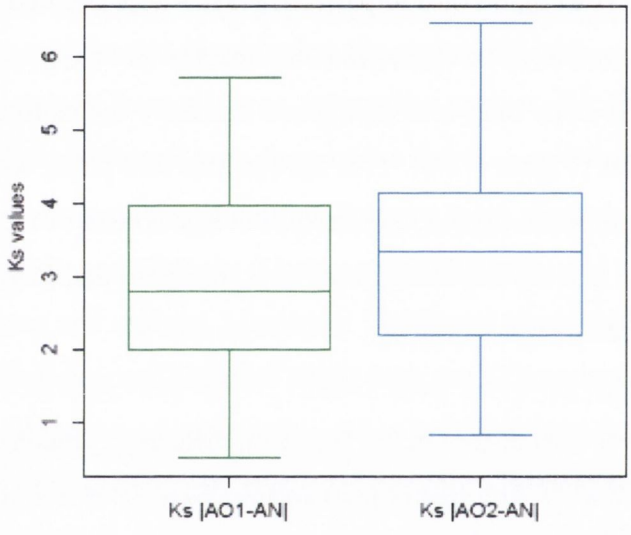


Figure 5.2 Boxplots of the distributions of synonymous nucleotide divergence (K_S) values between *A. nidulans* genes (*AN*) and *A. oryzae* gene copies *AO1* and *AO2*, for all genes in Set S_{AO}. The boxes show the median and interquartile values, with 95% of the data falling within the whiskers. K_S values were calculated using yn00 from the PAML package (Yang 1997a). K_S values for the *AO1* genes are significantly lower than for the corresponding *AO2* genes ($P < 10^{-4}$; Wilcoxon test).

5.4.3 No evidence for ancient WGD in *Aspergillus*

Another possible scenario that might explain the extra genes in *A. oryzae* is an ancient whole genome duplication (WGD) prior to the speciation of *A. oryzae*, *A. nidulans* and *A. fumigatus*, followed by parallel independent losses of most of the duplicated genes in the latter two species (Figure 5.3 Although this scenario might seem unlikely, it is a formal possibility so we searched for evidence of it.

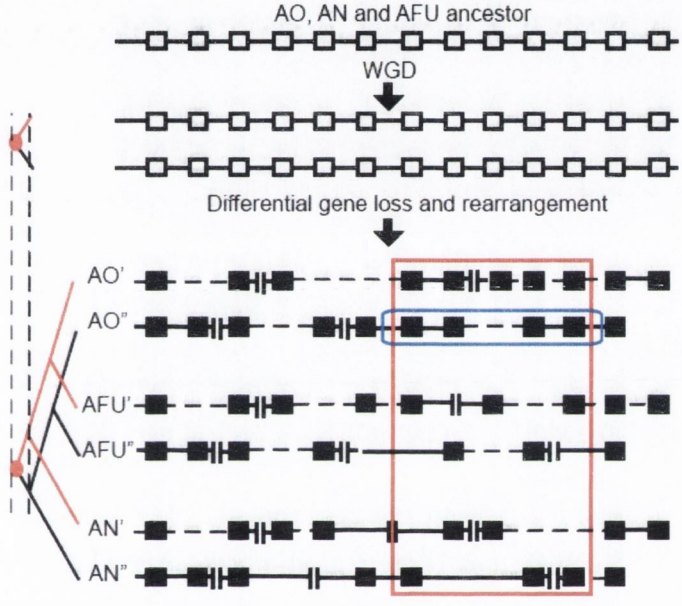


Figure 5.3 Model of a WGD hypothesis for the origin of extra genes in *A. oryzae*. A WGD in the common ancestor of the three *Aspergillus* species, followed by extensive gene losses in both *A. nidulans* and *A. fumigatus* could potentially account for the large number of genes in Set S_{AO}. In this schematic diagram black squares represent genes, double vertical lines represent chromosome breaks, and dashed horizontal lines represent gene deletions or intrachromosomal rearrangements. Our method of testing for WGD involves sliding a window (blue box) through the *A. oryzae* genome to find clusters of nearby genes that are enriched in genes from Set S_{AO} (red box). If a WGD occurred, the paralogs of the genes in the blue window are expected to be closer together on the sister *A. oryzae* genomic region than expected by chance.

The WGD events described so far in other genomes were detected using a number of different methods (reviewed in Van de Peer 2004). Patterns in the chromosomal locations of paralogs give by far the clearest signals in situations where a WGD is present in a sequenced genome, but no genome sequence from a closely-related unduplicated outgroup species is available (Wolfe and Shields 1997; Simillion et al. 2002; Dehal and Boore 2005; Aury et al. 2006; Jaillon et al. 2007). Immediately after a WGD, the genome consists of pairs of identical

chromosomes containing identical genes in the same order. This pattern will gradually become eroded by gene losses and chromosomal rearrangements, but we would expect that even an old WGD should leave a signal: the order of genes along a given chromosomal region should be correlated with the order of their paralogs along some other region in the genome. In other words, paralogs should not just be randomly distributed with regard to one another.

To compare the distribution of locations of paralogous genes in the *A. oryzae* genome to random expectations we defined a distance measure, *d*, of the extent to

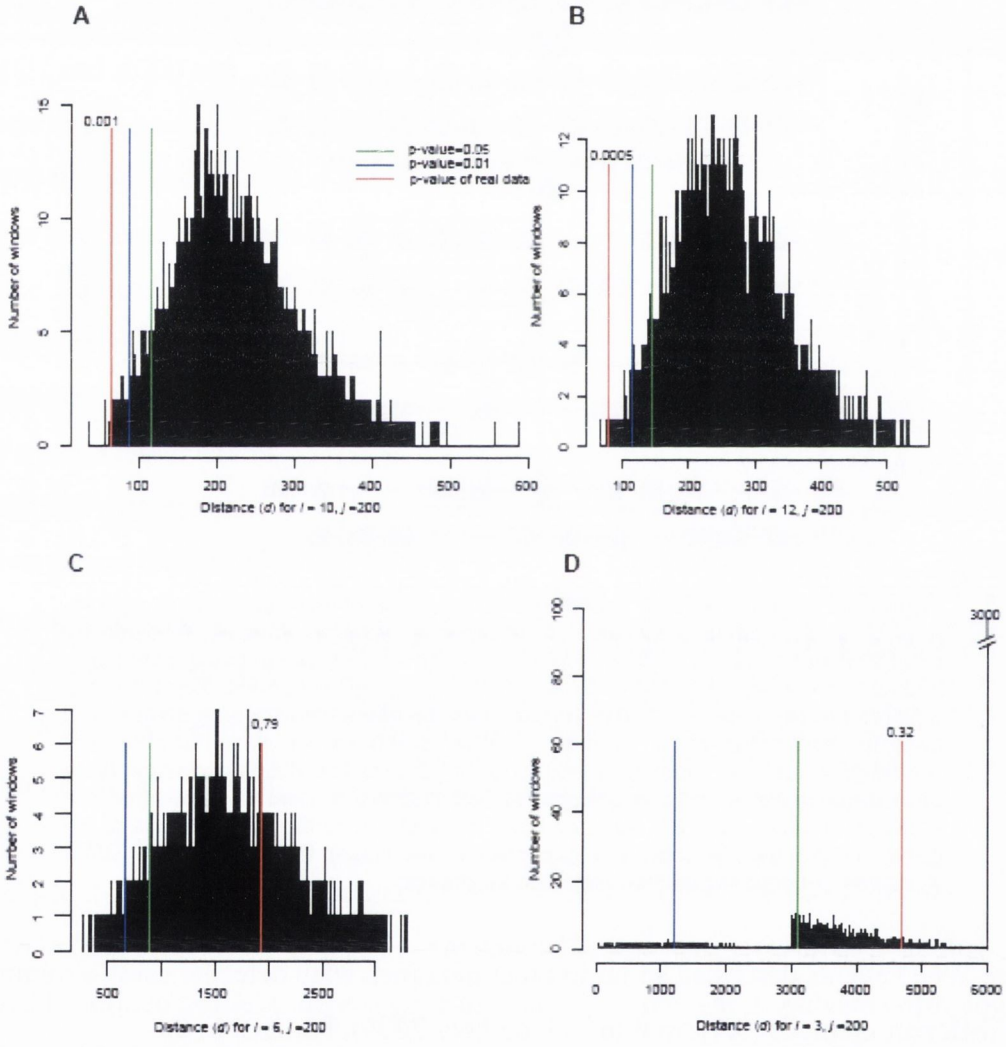


Figure 5.4 Testing for whole genome duplication in *A. oryzae* and *S. cerevisiae*. Black histograms show the distribution of distance scores (*d*) obtained in 10,000 simulations where paralog locations were randomized. Blue and green lines show the cutoff points for the most-clustered 1% and 5% of these data, respectively. Red lines show the observed *d* values for *S. cerevisiae* or *A. oryzae*. All the graphs are for windows of size *j* = 200. Panels A-C consider only windows where the paralogous genes are all on the same chromosome. (A) Results for *S. cerevisiae* for windows with *i* = 10 paralogs. (B) Results for *S. cerevisiae* for windows with *i* = 12 paralogs. (C) Results for *A. oryzae* for windows with *i* = 6 paralogs. (D) Results for *A. oryzae* for windows with *i* = 3 paralogs, including a penalty of 3000 for each interchromosomal rearrangement event.

which genes that are close together in a genome also have paralogs that are close together somewhere else in the genome (see Methods 5.3). We estimated the statistical significance of the observed distances in *A. oryzae* by comparing it to the distribution of distances obtained in computer simulations where gene locations were shuffled randomly (Figure 5.4).

We applied this method to both *S. cerevisiae* and *A. oryzae*. The method successfully detects a strong signal of non-random locations of paralogous gene pairs in *S. cerevisiae*, a species known to have undergone WGD (empirical $P < 10^{-3}$; Figure 5.4 A,B), but there is no significant signal of WGD in the *A. oryzae* genome ($P = 0.79$; Figure 5.4 C). This result not only argues against WGD but also against segmental duplication which would be expected to leave a similar, though weaker, signal. We also used a modified measure to permit consideration of situations where two genes on one chromosome have paralogs on two different chromosomes (see Methods 5.3) but again found no evidence of ancient WGD in *A. oryzae* (Figure 5.4 D).

In an independent analysis we carried out all-against-all BLASTP searches within the protein sets of each of *A. oryzae*, *A. nidulans* and *A. fumigatus*. We then made a dot matrix plot for each species showing the chromosomal locations of reciprocal-best-hit pairs within each genome. In previous studies on *S. cerevisiae* and plant species, approaches similar to this revealed strong diagonals corresponding to paralogous chromosomal segments that were formed by WGD (Wolfe and Shields 1997; Paterson et al. 2000; Jaillon et al. 2007). In contrast, inspection of our plots for *Aspergillus* species did not reveal any indication of WGD or segmental duplications (data not shown).

5.4.4 Can the excess of Topology S genes be explained by Horizontal Gene Transfer?

Taken together, the above results show that the excess of *A. oryzae* genes in Set S_{AO} cannot be explained satisfactorily by recent gene duplications in *A. oryzae*,

nor by an ancient WGD or segmental duplications in an *Aspergillus* ancestor. We therefore examined the third possibility of HGT into *A. oryzae*.

To detect a possible donor taxon, we first analyzed the phylogenetic trees of a random sample of genes from Set S_{AO}. This showed that the extra (*AO2*) copies of most genes are nested in the Ascomycota kingdom, so if HGT occurred the donor should also be in the Ascomycota. To provide a common reference point for phylogenetic reconstructions, we imposed the basidiomycete *Ustilago maydis* [whose genome is completely sequenced; 21], as the outgroup for all trees. This reduced our Set S_{AO} to a subset Q_{AO} of 122 genes for which an ortholog in

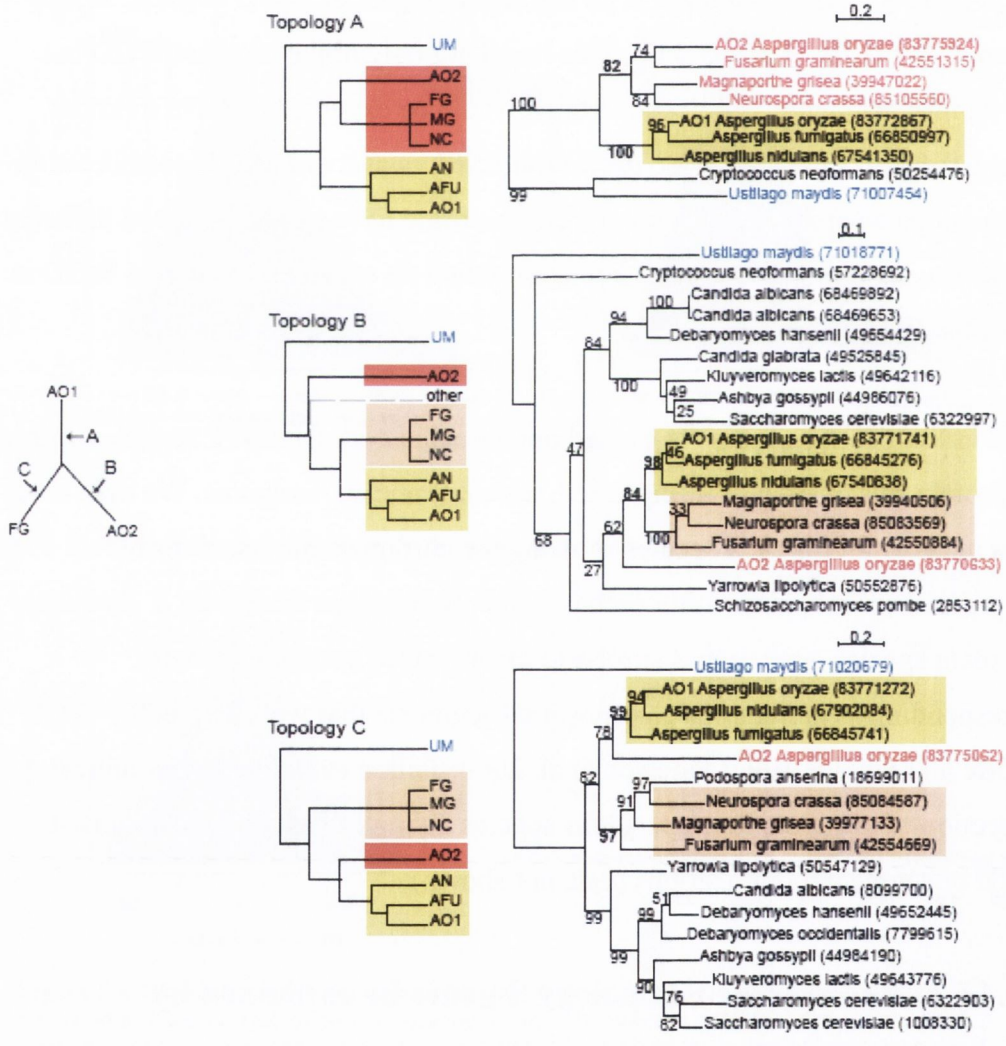


Figure 5.5 Topologies A, B and C for rooted trees. *Left*: The three topologies correspond to the three possible points for placement of a root (arrows labeled A, B, C) onto an unrooted tree consisting of a gene pair from *A. oryzae* and their homolog from a Sordariomycete such as *Fusarium graminearum*. *Center*: Rooted representations of the topologies. *Right*: real examples of *A. oryzae* genes classified in each of the three topologies. Numbers in parentheses are NCBI protein identifiers. Trees were constructed using PHYML with bootstrap percentages (form 500 replicates) shown. The main groups differentiating the topologies from each other are colored.

U. maydis could be identified unambiguously by a reciprocal-best-BLAST-hits (RBH) approach.

We constructed rooted phylogenetic trees for the 122 genes in Set Q_{AO} with their homologs in other fungi (primarily Sordariomycetes), and sorted them manually into three mutually exclusive topology types depending on the position of the *AO2* gene (Types A, B and C; Figure 5.5 ; Appendix A). In all three topologies the *AO1* gene of *A. oryzae* is clustered in a monophyletic group with its orthologs from *A. nidulans* and *A. fumigatus*, and in all three the species in the subphylum Sordariomycetes (*e.g., Neurospora crassa*, *Fusarium graminearum*, *Magnaporthe grisea*) form a monophyletic group. Type A trees are those where the second *A. oryzae* copy *AO2* clusters with Sordariomycete species to the exclusion of the genus *Aspergillus* (Figure 5.5 A). Type B trees have a topology in which *AO2* lies outside both the *Aspergillus* genus and the Sordariomycetes (Figure 5.5 B). Type C trees are those where *AO2* forms a sister group to the clade of *Aspergillus* orthologs that includes *AO1*, with Sordariomycete sequences outside this pair (Figure 5.5 C). Of these three topologies, only the Type A trees are directly suggestive of HGT because they implicate an identifiable donor lineage.

There are only three possible topologies for a phylogenetic tree that consists of an outgroup and three ingroup taxa. The A, B, and C trees correspond to these three topologies; the ingroup clades are *AO2*, the *Aspergillus* clade that includes *AO1*, and the Sordariomycete clade (Figure 5.5 , *left*). No other topologies can exist for our data unless the Sordariomycete or *Aspergillus* clades are broken up.

The sorting process for trees in Set Q_{AO} was done manually and with reference to the bootstrap support values for critical nodes on the tree. To keep the process simple we rejected trees from complex gene families or with poor bootstrap support, and we permitted minor deviations from the expected species phylogeny for the *Aspergillus* (*AO1*) and Sordariomycete clades. We also classified trees where the *AO2* gene was placed within the Sordariomycetes, rather than as sister to Sordariomycetes, as Type A. These additional restrictions left only 35 classifiable trees from Set Q_{AO}, of which 12 were Type A, 9 were Type B, and 14 were Type C (Table 5.1 and Table 5.2). The fact that we find approximately

equal numbers of trees of the three possible types is not consistent with the hypothesis that all *AO2* genes originated by HGT from the same donor. Instead, the result suggests that the *AO2* genes are heterogeneous in terms of their origins: if they all arose by HGT then the HGT events involved multiple donors, or if they all arose by gene duplication then those duplications occurred at multiple times.

Topology	Number of trees		
	A. oryzae (Set Q_{AO})	*A. nidulans* (Set Q_{AN})	*A. fumigatus* (Set Q_{AFU})
A	12	0	0
B	9	9	2
C	14	2	4

Table 5.1 Numbers of phylogenetic trees with Topologies A, B and C obtained from genes in Sets Q_{AO}, Q_{AN} and Q_{AFU}.

For comparison, we made similar analyses in *A. nidulans* and *A. fumigatus*. We compiled two other sets of genes that are analogous to Set Q_{AO} but have duplications only in *A. nidulans* (Set Q_{AN}, 40 genes), or only in *A. fumigatus* (Set Q_{AF}, 28 genes), and that give phylogenetic trees that place the second gene (*AN2* or *AFU2*) outside a monophyletic group containing one gene from each of the three *Aspergillus* species. As expected, these gene sets are smaller than the corresponding set with duplication in *A. oryzae*. After phylogenetic tree construction we find that sets Q_{AN} and Q_{AFU} both present an absence of Topology A trees (Table 5.1). Although the numbers of classified trees are small (11 in Set Q_{AN} and 6 in Set Q_{AFU}) this deficit contrasts with the number of Topology A trees seen in the *A. oryzae* set. Topology A trees are the only ones where a putative HGT donor lineage can be identified (Figure 5.5 so it is striking that trees with this topology are found only in *A. oryzae*. If HGT from Sordariomycetes to *A. oryzae* accounts for the type A topologies, we can estimate that it maximally

accounts for one-third of the extra genes in *A. oryzae* (that is, by extrapolation from the 12/35 proportion found in the loci that were classifiable).

It is also notable that among the 12 *A. oryzae* gene pairs that give Topology A trees, the trees for 10 of these pairs place the *AO2* gene on a branch *within* the Sordariomycetes as opposed to lying sister to the Sordariomycetes. An example is the Topology A gene in Figure 5, where the *AO2* sequence (NCBI identifier 83775924) clusters specifically with *F. graminearum*. Of the 10 genes that

Table 5.2 Trees classified as Types A, B and C in each *Aspergillus* species. Trees were constructed using PHYML as described in *Methods*. In each tree, the sequences identified as *AO1* and *AO2* (for duplications in *A. oryzae*), *AN1* and *AN2* (for duplications in *A. nidulans*), or *AFU1* and *AFU2* (for duplications in *A. fumigatus*) are labeled. NCBI identifier (GI) numbers for each sequence are shown.

AO1	AO2	Protein domains present
		Topology A
AO090701000472	AO090103000225	Isopenicillin N synthase and related dioxygenases family
AO090120000414	AO090010000135	Iron transporter
AO090003000721	AO090009000136	Homoserine dehydrogenase, NAD binding domain
AO090001000581	AO090010000132	Serine/threonine kinase TIP30/CC3. Predicted nucleoside-diphosphate-sugar epimerase
AO090102000393	AO090003000572	Fatty acyl-CoA elongation enzyme
AO090009000593	AO090103000299	Metal-dependent amidase/aminoacylase/carboxypeptidase
AO090012000848	AO090009000109	Haloacid dehalogenase-like hydrolase
AO090023000127	AO090012000890	Dihydroxy-acid dehydratase
AO090003000571	AO090166000076	Acetohydroxyacid synthase (AHAS) subfamily
AO090206000073	AO090011000052	Protease-associated (PA) domain
AO090011000374	AO090003000277	Short chain dehydrogenase
AO090102000131	AO090012000011	Conserved protein of unknown function
		Topology B
AO090102000148	AO090010000406	Mitochondrial carrier protein
AO090102000626	AO090113000140	Mitochondrial carrier protein
AO090113000163	AO090026000246	X-Pro dipeptidyl-peptidase (S15 family)
AO090003000972	AO090023000534	Protease-associated (PA) domain
AO090012000452	AO090012000771	Conserved protein of unknown function
AO090038000292	AO090012000674	Dehydrogenase
AO090701000122	AO090012000618	Phosphodiesterase / nucleotide, GPI anchor synthesis protein
AO090001000454	AO090120000159	Mitochondrial carrier protein
AO090020000590	AO090701000520	2-Nitropropane dioxygenase
		Topology C
AO090010000480	AO090124000011	Actin depolymerisation factor/cofilin
AO090701000373	AO090308000002	Zn-dependent alcohol dehydrogenase
AO090026000284	AO090023000874	Tryptophan synthase, beta chain
AO090005001147	AO090701000660	Putative NADP-dependent oxidoreductase
AO090026000661	AO090166000064	Heavy metal exporter HMT1, ABC superfamily
AO090120000229	AO090010000496	Histidine kinase-like ATPase, HAMP domain
AO090102000109	AO090012000014	Unknown
AO090701000582	AO090103000289	Serine racemase
AO090003000922	AO090124000052	Ca2+/H+ antiporter VCX1 and related proteins
AO090206000105	AO090010000720	Ribonuclease T2
AO090003000627	AO090001000179	RTA1 like protein
AO090011000825	AO090138000089	Mitochondrial endonuclease
AO090701000555	AO090103000285	Enoyl-CoA isomerase
AO090102000343	AO090011000734	Na+/K+ transporter

showed this property, five clustered with *F. graminearum*, three with *M. grisea* and two with *N. crassa*. A single consistent candidate donor lineage is again lacking, even within the Sordariomycetes.

5.4.5 Functions and physical clustering of AO2 genes with Topology A within the *A. oryzae* genome

The set of 12 *A. oryzae* genes with Topology A is enriched in particular functions, most notably the hydrolytic traits that are characteristic of *A. oryzae* (Table 5.2).

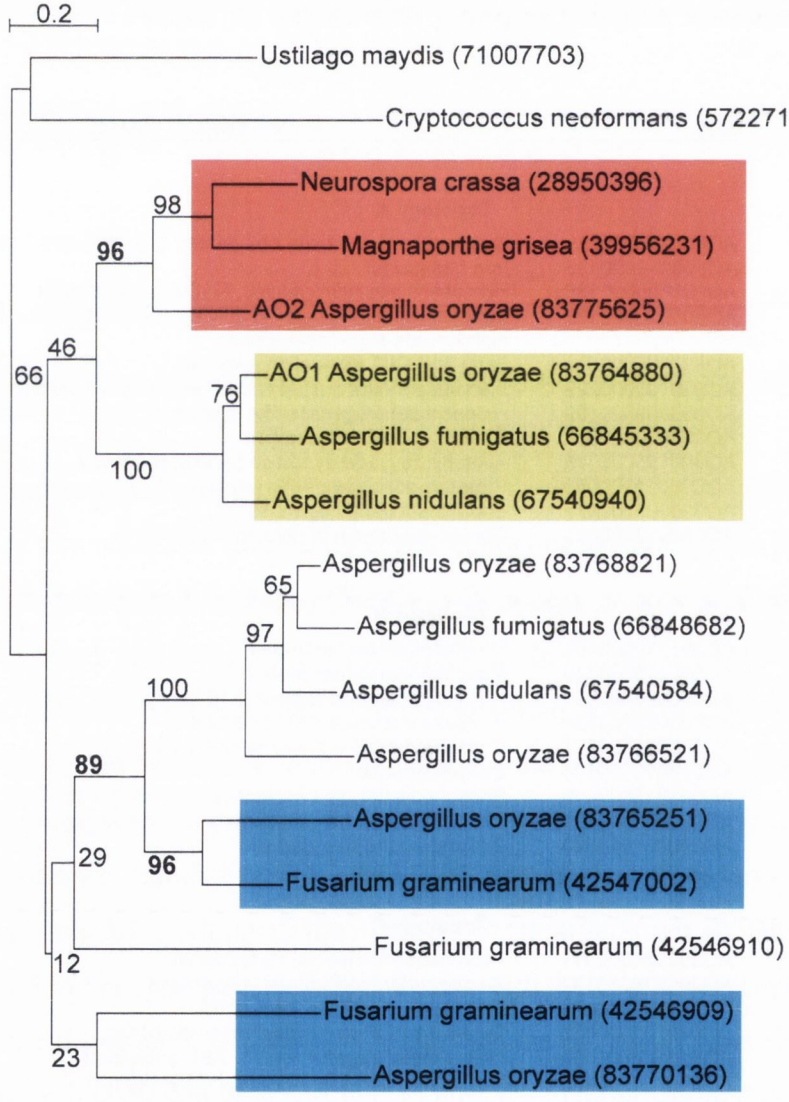


Figure 5.6 A case of family expansion due to apparent HGT. The gene family is a member of the metal dependent hydrolytic enzyme superfamily (amidases, aminoacyclases, and carboxypeptidases). The *A. oryzae* genes *AO1* and *AO2* originally identified in our automated search are labeled. Four of the six *A. oryzae* genes (83775625, 83765251, 83766521, 83770136) in this family lack orthologs in *A. nidulans* and *A. fumigatus*, and for the first two of these there is strong bootstrap support for HGT from Sordariomycete. Of these four genes, three show Topology A and one (83766521) shows Topology C relative to their nearest Sordariomycete homologs.

This species breaks down starch into sugars, and proteins into amino acids, by producing vast amounts of hydrolytic enzymes. The enrichment of some families in these classes among duplicated *A. oryzae* genes is striking, as previously also noted by Machida *et al.* (Machida et al. 2005). For example, gene AO090103000299 is a carboxypeptidase that our method identified as an *AO2* gene giving Topology A. Further exploration of this gene family (Figure 5.6 shows that it is a member of a six-gene family in *A. oryzae*, only two of which have orthologs in *A, nidulans* and *A. fumigatus*. Indeed this family consists of only two genes in the latter two species. Of the four other *A. oryzae* genes in the family, three cluster with Sordariomycete genes in the phylogenetic tree (two with different *F. graminearum* genes, and one (the gene AO090103000299 originally identified as an *AO2* candidate) with a *M. grisea-N. crassa* pair. In this *A. oryzae* gene family, only two genes were inherited vertically and are shared with other *Aspergillus* species, two (possibly three) originated by HGT from Sordariomycetes, and the sixth gene has an undetermined origin. Thus, HGT appears to have played a major role in the expansion of this carboxypeptidase family in *A. oryzae*.

Comparison to the genome sequence of *A. flavus* (G. Payne *et al.*, unpublished), which has a recent and clonal relationship to *A. oryzae* (Geiser et al. 1998), shows that many of the *AO2* genes have orthologs in that species. Therefore, the *AO2* genes were gained by *A. oryzae* after it diverged from *A. fumigatus* but before the split between *A. flavus* and *A. oryzae*.

If foreign DNA enters a recipient genome in a single event, the foreign genes might be expected to exhibit some clustering in the recipient genome. Conversely if they arrived in independent events, or if the hypothesis of HGT is incorrect, they should exhibit random genomic locations. If an HGT event is very recent, the transferred genes may also still be clustered in the donor genome.

We applied the method of Lee and Sonnhammer (Lee and Sonnhammer 2003) to test whether the locations of the *AO2* genes that may have undergone HGT from Sordariomycetes show significant physical clustering in the *A. oryzae* genome. For the 12 *AO2* genes with tree topology A (Table 5.2), the clustering score is 56,

as compared to a mean score of 21.7 for 100,000 replicates with randomized data (empirical $P = 0.02$). In contrast, the *AO1* partners of these genes do not show physical clustering (score = 6.6; $P = 0.95$), nor did we find significant clustering in comparisons of the *A. oryzae AO2* genes that gave B, C, or both B and C topologies ($P = 0.19$; $P = 0.06$ and $P = 0.19$ respectively). This result again highlights the distinctive nature of the *AO2* genes with tree topology A, in terms of both their phylogenetic and genomic placement. The physical clustering of the Topology A *AO2* genes is indicative of either a one-time transfer event, or a preference for integrating new genes into the genome at particular sites. We did not detect significant clustering in the *F. graminearum* genome of homologs of the *AO2* genes that form branches specifically with *F. graminearum* (though the dataset is small; only 5 genes).

5.5 Discussion

A frustrating feature of our study was the extent of data loss that we encountered as the analysis progressed. The *A. oryzae* genome contains more than one thousand extra genes as compared to *A. nidulans* and *A. fumigatus*, but when we attempted to use phylogenetic methods to determine the origin of these genes the dataset collapsed to only 35 informative loci (Table 5.1 One major cause of the reduction of the dataset is that we were only able to work with genes that are duplicated in *A. oryzae*. The reason for this is pragmatic: we cannot actually identify which, out of the 12,074 genes in *A. oryzae*, are the 'extra' ones, except in the 456 cases where the *AO1* gene was retained too (Set S_{AO}). A second problem was the lack of identifiable basidiomycete outgroup sequences for many of the genes that are duplicated in *A. oryzae*. It is difficult to speculate why these sequences do not exist, but one possibility is that the set of genes that has been added to the *A. oryzae* genome tends to be derived from the fastest-evolving, and hence ascomycete-specific, portion of the fungal genome (Cai et al. 2006; Nishida 2006).

We were able to rule out annotation error and WGD as possible sources of the extra genes in *A. oryzae* relative to other *Aspergillus* species. Our phylogenetic

analysis also rules out species-specific gene duplication events in *A. oryzae* as the source of many (456) of these genes. These results leave only two viable possibilities for the origin of the extra genes: HGT, or older gene duplications (in the common ancestor of the three *Aspergillus* species) followed by independent parallel losses in *A. nidulans* and *A. fumigatus*. The second of these possibilities is unparsimonious – we do not see why the level of retention of ancient duplicated genes should have been so much higher in *A. oryzae* than in the other two *Aspergillus* species – and it is specifically ruled out (barring phylogenetic tree error) in the case of the 12 loci with Topology A (Figure 5.5 and Table 5.2). This result seems to leave HGT as the only viable option for the source of the increased gene content of *A. oryzae*, which makes it all the more puzzling that we were unable to identify a consistent donor lineage. Instead, if we accept that all the tree topologies are correct, we must postulate that *A. oryzae* has received genes from at least five different donor lineages: one lineage that diverged prior to the Eurotiomycetes/Sordariomycetes split (giving Topology B trees), one lineage that is in Eurotiomycetes but outside *Aspergillus* (Topology C), and three separate Sordariomycete lineages that have specific affinities to *F. graminearum*, *N. crassa* and *M. grisea* respectively (accounting for the heterogeneity within the Topology A trees). We can speculate that some aspect of the lifestyle or physiology of *A. oryzae* (and *A. flavus*) could make it more able to take up foreign DNA by HGT than other *Aspergillus* species.

Many previous published examples of HGT have involved genes that were transferred between distantly related kingdoms, such as from bacteria to eukaryotes (Garcia-Vallve et al. 2000; Nixon et al. 2002; Andersson et al. 2003; Hall et al. 2005; Frigaard et al. 2006; Ricard et al. 2006). In many of these cases the gained genes were inferred to have changed the lifestyle of the recipient species markedly. In our case we are studying a more local exchange between reasonably closely related genomes that already share quite similar gene contents and functions. Our method of analysis necessitates that all the genes we identify as putative gains by HGT result in gene family expansion rather than in the acquisition of new biochemical functions. It is tempting to speculate that HGT helped *A. oryzae* adapt to some new evolutionary niche, but based on our analysis we can only assert that HGT enhanced the hydrolytic and fermentation abilities of

this species. We can infer that, because these foreign genes were taken up and maintained in the genome, they contributed in some way to increasing *A. oryzae*'s evolutionary fitness, but we do not know exactly what that contribution was.

5.6 Acknowledgments

We thank Gary Payne and members of his lab for sharing *A. flavus* data, and are grateful to Gavin Conant, Kevin Byrne and Meg Woolfit for discussion.

Chapter 6

Effects of domestication on genome evolution in *Aspergillus flavus* and *A. oryzae*

The work described in this short chapter will form part of a larger comparison of the *A. flavus* and *A. oryzae* genomes, which is being coordinated by Prof. Gary Payne (North Carolina State University).

6.1 Abstract

How domestication affects and directs the evolution of a given population is fascinating. It demonstrates the power of human beings to alter, by artificial selection, the DNA sequence of the genome of another species. Further, it shows how relatively short evolutionary time scales can split one species into two, or can create new species. In this chapter we investigate the genomic differences between *Aspergillus flavus* and *A. oryzae*, which originated from *A. flavus* by domestication. *A. flavus* is an opportunistic plant and animal pathogen that produces aflatoxin, a potent carcinogen. In contrast, *A. oryzae* is used extensively in the food industry and thus is regarded as safe. Whole genome sequence data for these two species allowed us to closely examine their genetic relationship. DNA sequence comparison shows a significantly higher non-synonymous substitution rate (*dN*) in *A. oryzae* than in *A. flavus*, indicating different selection pressures for these genes in the two fungi. We show that different sets of genes are accelerating in each species, and identify rapidly-evolving genes that are important for the lifestyle of each species.

6.2 Introduction

Domestication can result in massive changes of phenotype, as illustrated most dramatically by the huge range of morphological variation seen among dog breeds (Darwin 1859; Lindblad-Toh et al. 2005), or by the vast differences between domesticated maize plants and their teosinte-like ancestors (Doebley et al. 2006). These phenotypic changes are the result of artificial selection by human breeders during the past 10,000 years or so. Underlying these phenotypic changes are changes to the genome.

Some studies have shown that domestication can have dramatic effects on genomes. For example, mitochondrial genes of domesticated dogs have significantly accelerated in sequence evolution compared to their orthologs in wolves (Bjornerfeldt et al. 2006). This acceleration has been attributed to relaxation of constraint on the dog genome. Indeed, the first domesticated wolves did not need to practice many of their usual daily tasks; their lifestyle became very different to what it was in the wild. The authors hypothesized that changes in the living conditions of dogs as a result of domestication resulted in the release of selective constraint, allowing a faster accumulation of functional genetic diversity in many of dog genes. Similarly, comparison of the genome of a "domesticated" strain (S288c) of the yeast *S. cerevisiae* that has been maintained in rich media in laboratories for the past 80 years, to a "wild" strain (YJM789) recently isolated from a clinical infection, also showed accelerated evolution of protein sequences in the domesticated strain, presumably due to relaxation of selective constraints (Gu et al. 2005).

The *Aspergillus* species *A. flavus* and *A. oryzae* are morphologically very similar (Changa et al. 2006). Indeed, little or no differences are observed at the genome level. However *A. flavus* is an important producer of aflatoxin, and is a potent opportunistic pathogen. On the other hand *A. oryzae* is used in the food industry. *A. oryzae* is incapable of synthesizing aflatoxin due to multiple molecular lesions in its aflatoxin gene cluster: a deletion in the *aflT* region, a frameshift mutation in *norA*, a base pair substitution in *verA*, and two substitutions in the binding sites of

the *aflR* promoter (Tominaga et al. 2006). The important differences in their lifestyles led us to examine the effects of domestication on the evolution of *A. oryzae*.

6.3 Methods

The *A. nidulans* genome sequence was downloaded from the Broad Institute at: http://www.broad.mit.edu/annotation/genome/aspergillus_nidulans/Downloads.html/direct;jsessionid=3EC5C666C94729B02CD38964EA988E1B. The *A. oryzae* genome was downloaded from National Center for Biotechnology Information (NCBI). *A. flavus* was provided by Doctor Gary Payne.

To find orthologs, we identified three-way mutual best BLASTP hits among *A. oryzae*, *A. flavus* and *A. nidulans*. This method resulted in 6057 sets of putative orthologs among the three species. Each trio of proteins was aligned using ClustalW (Thompson et al. 1994). Poorly aligned regions were removed using Gblocks (Castresana 2000). We used codeml from the PAML package (Yang 1997b) to estimate *dN* and *dS* values for the three branches. Alignments with exceptionally high *dN* values were examined by eye, and 159 trios were eliminated due to poor alignment. The sets of *A. flavus*-accelerated and *A. oryzae*-accelerated genes were identified as the 5% of genes at each end of the distribution of the quantity *dN*(AO)-*dN*(AFL) as shown in Figure 6.1 b; this is an *ad hoc* cutoff that does not make any assumption about the shape of the distribution. We used *dN* in preference to the *dN/dS* ratio because the low divergence between the species means that for some genes *dS* is zero, and because for a set of orthologous genes (as here) we expect that all gene pairs diverged simultaneously so there is no need to correct the *dN* values for divergence time.

Gene Ontology (GO) terms were imported to *A. oryzae* and *A. flavus* genes based on the terms assigned to their orthologs in other species. Primary GO terms were imported from *A. fumigatus* (Nierman et al. 2005), and GO-slim terms were imported from both *A. fumigatus* and *S. cerevisiae*. The significance of GO term

usage in gene sets was estimated using Fisher's exact test for primary GO terms, and a chi-square test for GO-slim terms.

Estimating divergence dates between fungi is difficult because few fossils or archaeological data exist with which to calibrate molecular clocks. The divergence time between *A. oryzae* and *A. flavus* is clearly very recent because the median *dS* between these species is only 0.00096. This is about 13 times lower than the average *dS* between human and chimpanzee (0.0124), two primates that diverged 4.6–6.2 million years ago (Chen and Li 2001). If we apply the primate molecular clock for synonymous substitutions to the fungi, the *A. oryzae-A. flavus* divergence time is estimated as 360,000–480,000 years ago. This calculation is obviously rather crude, but it gives an idea of the order of magnitude of the divergence time. The median *dS* value between *A. oryzae* and *A. flavus* is also 10 times lower than the mean *dS* reported between two strains of *S. cerevisiae* (S288c and YJM789) (Gu et al. 2005).

6.4 Results and Discussion

It is widely agreed that *A. oryzae* is a domesticated variety of *A. flavus* that has lost its potential for pathogenicity. There are very few differences between the species at the genomic level in terms of gene content or gene order. This led us to examine the question of whether hallmarks of *A. oryzae*'s domestication can be detected in its genome. Previous analyses have shown that domestication may involve a relaxation of selective constraint, and subsequent change in the rate and pattern of evolution in domesticated species (Gu et al. 2005; Bjornerfeldt et al. 2006; Doebley et al. 2006).

We used a maximum-likelihood (ML) approach to estimate the levels of synonymous (*dS*) and non-synonymous (*dN*) nucleotide substitution in genes along each branch of the gene tree (Figure 6.1 a), using *A. nidulans* as the outgroup, for 6057 genes. We found that *dN* is significantly higher on the *A. oryzae* branch than on the *A. flavus* branch ($P < 10^{-15}$ by Wilcoxon signed ranks test), whereas *dS* does not present any significant difference between the species

(medians 7×10^{-6} and 7×10^{-6} respectively; P = 0.34). The asymmetry of nonsynonymous rates is illustrated by the finding that *dN* is higher in *A. oryzae* than in *A. flavus* for 3269 genes, whereas the converse is true for only 2213 genes (Figure 6.1 b). These results point towards a relaxation of constraint on *A. oryzae*'s genome. This relaxation is probably due to one or more bottlenecks that the *A. oryzae* population went through during its evolution.

We estimate roughly that the time of divergence between the *A. oryzae* and *A. flavus* species is on the order of 400,000 years (see Methods 6.3). Therefore, although *A. oryzae* is likely to have gone through a genetic bottleneck when it was domesticated by humans a few thousand years ago, the period of time during which humans have influenced its evolution is only a small fraction of its divergence time from *A. flavus*.

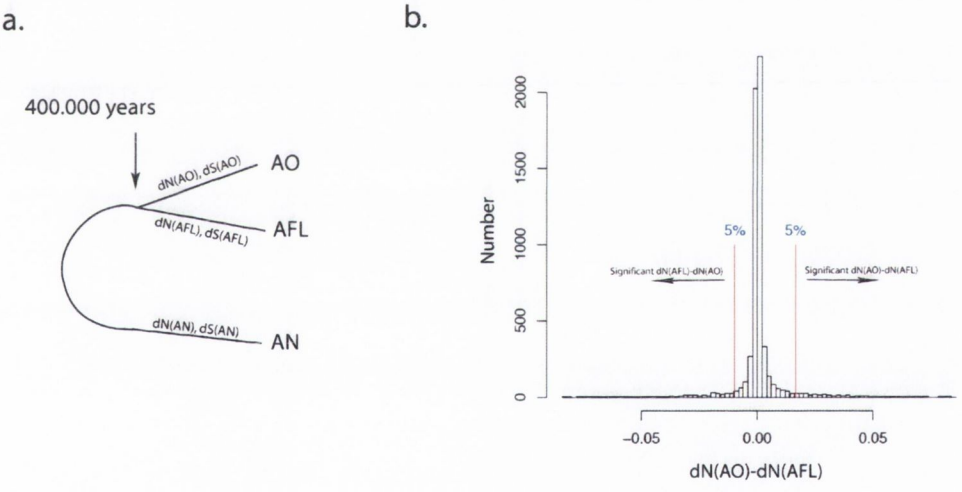


Figure 6.1 (a) Tree relating *A. oryzae*, *A. flavus* and *A. nidulans*. The *dN* and *dS* quantities on each branch refer to divergences from the central node of the unrooted tree. (b) Histogram of the differences in *dN* between the *A. oryzae* and *A. flavus* branches. The red lines indicate the cutoffs for the 5% of genes at each end of the distribution. Note that the center of the overall distribution lies to the right of the zero point, indicating faster evolution on the *A. oryzae* lineage than the *A. flavus* lineage.

Besides a relaxation of constraint caused by population size reduction, we might also expect that *A. oryzae*'s dramatic change of niche (from plant pathogen to food industry solid-state fermenter) could have substantially changed the selective pressures acting on many of its genes, resulting in turn in elevated levels of substitution. What is important for one species may not be so important for the other. In addition, human intervention may have specifically selected for certain alleles in *A. oryzae*, thus affecting particular genes. To investigate both of these

possibilities, we examined the sets of genes with the largest differences between their nonsynonymous values on the *A. oryzae* and *A. flavus* branches. We arbitrarily defined these sets ("*A. oryzae*-accelerated" and "*A. flavus*-accelerated") as the two tails of the distribution shown in Figure 6.1 b, each of which contains 5% of the data (303 genes each).

We used Gene Ontology analysis to look for differences between each of these two sets of genes and the complete set of *A. oryzae-A. flavus* orthologs. We initially used GO-slim terms, but very few genes in the sets accelerated in *A. flavus* and *A. oryzae* have been associated to these terms, which made difficult to test for significance. We therefore switched to analyses using the full set of GO terms. We find that cytochromes are over-represented in the *A. flavus*-accelerated list. Of the 47 cytochrome genes in the set of orthologs, six are in the set accelerated in *A. flavus* (P = 0.02545). Five of these genes are cytochrome P450 oxidoreductases, implicated in toxin production and detoxification, consistent with *A. flavus*'s pathogenic lifestyle. Only one cytochrome P450 is in *A. oryzae*-accelerated set; this gene is part of the aflatoxin cluster, which is nonfunctional in *A. oryzae* and is therefore not expected to be under any constraint in *A. oryzae*. Four other genes in the aflatoxin cluster are also found in the *A. oryzae*-accelerated set: *moxY*, *norA*, *adhA* and *avfA*. The presence of such a number of genes from the same cluster confirm the expected relaxation of evolutionary constraints on the aflatoxin cluster in *A. oryzae*, and validate the approach we have taken to identify genes under differential constraint in the two species.

In the *A. flavus*-accelerated set we also find a significant overrepresentation of membrane transporters protein from the Major Facilitator Superfamily (MFS) (11 genes out of 110 in total; P = 0.004), and significant overrepresentation of hydrolase genes (14 out of 163, P = 0.04). A significant number of the latter are glycosyl hydrolases (7 out of 50; P = 0.01). Three glycosyl hydrolases families are represented (Davies and Henrissat 1995; Henrissat and Davies 1997). Family 1 (GH1) has been implicated in a diverse processes of physiological importance, such as biomass conversion in microorganisms (1993) and activation of defense compounds (Poulton 1990; Duroux et al. 1998). Family 3 (GH3) is required for optimal growth on cellobiosides (Faure et al. 2001). These enzymes also modulate

the biological activities of different β-glucosides, such as antibiotics (Quiros et al. 1998) or saponins (Morrissey and Osbourn 1999). Family 17 (GH17), found in the cell wall matrix of plants and fungi (Hrmova et al. 1997), is implicated in diverse processes one of which is activation of defense compounds.

Although they are not significantly over-represented, it is worth mentioning that three sugar transporters (out of 37) are present in the *A. flavus*-accelerated set. Together, sugar transporters, hydrolases and MFS are often connected to a saprophytic or/and plant pathogen lifestyle. *A. flavus* is a saprophyte and occasionally an opportunistic pathogen. It is difficult to know why these genes are accelerating in *A. flavus*, but one reason might be that *A. flavus* is shifting from a saprophyte lifestyle to a more pathogenic one; indeed, *A. flavus* is not regarded as a good saprophyte (personnel communication from Gary Payne).

Other genes of interest in the *A. flavus*-accelerated list include the vacuolar heavy metal ABC transporter (Hmt1), whose ortholog in *Schizosaccharomyces pombe* has been shown to be important in detoxification of xenobiotics (Ortiz et al. 1995); the *meaA* ammonium transporter, whose *A. nidulans* ortholog is required for optimal growth on ammonium as the sole nitrogen source (Monahan et al. 2002); and the ortholog of the major *A. fumigatus* allergen Asp F2, which is associated with allergic bronchopulmonary aspergillosis (Banerjee et al. 1998).

The set of *A. oryzae*-accelerated genes is noticeably enriched in transcriptional regulators (2 out of 7 genes with this phrase in their description; $P = 0.04$) and zinc finger proteins (5 out of 14; $P = 0.0004$) as compared to the complete set of orthologs. However, the larger GO term "transcription factors" is not significantly overrepresented in the *A. oryzae*-accelerated set.

These comparisons of accelerated gene sets reflect how a few thousand years of separate evolution in very different environments has had significant effects on both genomes. Genes that are implicated in detoxification and interaction with the exterior are significantly over represented in the *A. flavus*-accelerated list but not in *A. oryzae*. On the other hand, the set of genes whose rate of protein sequence

evolution has accelerated in *A. oryzae* includes regulatory genes and genes of the aflatoxin cluster.

Chapter 7

Genomic Islands in the Pathogenic Filamentous Fungus *Aspergillus fumigatus*

7.1 Publication

The research described in this chapter has been published: Fedorova ND, **Khaldi N**, Joardar VS, Maiti R, Amedeo P, Anderson MJ, Crabtree J, Silva JC, Badger JH, Albarraq A, Angiuoli S, Bussey H, Bowyer P, Cotty PJ, Dyer PS, Egan A, Galens K, Fraser-Liggett CM, Haas BJ, Inman JM, Kent R, Lemieux S, Malavazi I, Orvis J, Roemer T, Ronning CM, Sundaram JP, Sutton G, Turner G, Venter JC, White OR, Whitty BR, Youngman P, Wolfe KH, Goldman GH, Wortman JR, Jiang B, Denning DW, Nierman WC, *PLoS Genetics* 4:e1000046, 2008).

The author contributed to writing the manuscript, and was responsible for the following three sections and figures:

- The phylogeny of the Aspergilli species and created Figure 7.2 .
- Contributed to Section 7.5.2 on 'Highly Variable Loci in *A. fumigatus*', and created Figure 7.4 .
- Responsible for Section 7.5.4.4 on 'Evolutionary Origins of Lineage-Specific Genes' and created Figure 7.7 .

7.2 Abstract

We present the genome sequences of a new clinical isolate of the important human pathogen, *Aspergillus fumigatus*, A1163, and two closely related but rarely pathogenic species, *Neosartorya fischeri* NRRL181 and *Aspergillus clavatus* NRRL1. Comparative genomic analysis of A1163 with the recently sequenced *A. fumigatus* isolate Af293 has identified core, variable and up to 2% unique genes in each genome. While the core genes are 99.8% identical at the nucleotide level, identity for variable genes can be as low 40%. The most divergent loci appear to contain heterokaryon incompatibility (*het*) genes associated with fungal programmed cell death such as developmental regulator *rosA*. Cross-species comparison has revealed that 8.5%, 13.5% and 12.6%, respectively, of *A. fumigatus*, *N. fischeri* and *A. clavatus* genes are species-specific. These genes are significantly smaller in size than core genes, contain fewer exons and exhibit a subtelomeric bias. Most of them cluster together in 13 chromosomal islands, which are enriched for pseudogenes, transposons and other repetitive elements. At least 20% of *A. fumigatus*-specific genes appear to be functional and involved in carbohydrate and chitin catabolism, transport, detoxification, secondary metabolism and other functions that may facilitate the adaptation to heterogeneous environments such as soil or a mammalian host. Contrary to what was suggested previously, their origin cannot be attributed to horizontal gene transfer (HGT), but instead is likely to involve duplication, diversification and differential gene loss (DDL). The role of duplication in the origin of lineage-specific genes is further underlined by the discovery of genomic islands that seem to function as designated "gene dumps" and, perhaps, simultaneously, as "gene factories".

-

7.3 Introduction

Aspergillus fumigatus is exceptional amongst the aspergilli in being both a primary and opportunistic pathogen as well as a major allergen associated with severe asthma and sinusitis (Denning 1998; Latge 1999). It was first reported to cause opportunistic invasive infection about 50 years ago (Rankin 1953). In immunocompromised patients, mycelial growth can proliferate throughout

pulmonary or other tissues causing invasive aspergillosis. For these patients, the incidence of invasive aspergillosis can be as high as 50% and the mortality rate is often 50%, even with antifungal treatment. Since the late 1800's (Denning 1998), *A. fumigatus* has been demonstrated to be a primary pathogen of the airways, sinuses, lungs, damaged skin and subcutaneous tissues. For example, it can cause post-operative infection in all human organs (Pasqualotto and Denning 2006). In most cases diagnosis remains problematic and can compromise effective medical treatment.

A. fumigatus is thought to possess particular metabolic capabilities and genetic determinants that allow it to initiate and establish an *in vivo* infection. This conclusion is supported by the observation that the majority of invasive aspergillosis disease is caused by *A. fumigatus*, even though its conidia comprise only a small percentage of the total conidia found in air-sampling studies (Richardson et al. 2000). While the interaction of *A. fumigatus* spores with the human respiratory mucosa is understood to an extent, the basic biology of the organism has until recently received little attention.

Recently we presented the genomic sequence of *A. fumigatus* strain Af293 (FGSC A1100) (Nierman et al. 2005) isolated from a neutropenic patient, who died from invasive aspergillosis (Payne et al. 2006). Its comparison with the genomes of two distantly related species, *Aspergillus nidulans* and *Aspergillus oryzae,* has led to many unexpected discoveries, including the possibility of a hidden sexual cycle in *A. fumigatus* and *A. oryzae*, and the detection of remarkable genetic variability of this genus (Galagan et al. 2005; Machida et al. 2005). Although members of the same genus, these three species are approximately as evolutionarily distant from each other at the molecular level as humans and fish (Figure 7.1 and Figure 7.2) (Taylor and Berbee 2006). This significant phylogenetic distance has hindered some aspects of comparative genomic analysis of the aspergilli such as identification of the genetic traits responsible for differences in virulence as well as in sexual and physiological properties.

To maximize the resolving power of whole-genome comparative analysis, we selected the environmental type strains of a very closely related sexual species,

Neosartorya fischeri NRRL181 (*A. fischerianus*), and a more distantly related asexual species, *A. clavatus* NRRL1, for complete sequencing. These three species are referred to here as the Affc lineage for *A. fumigatus*, *N. fischeri*, and *A. clavatus* (Figure 7.2). In contrast to *A. fumigatus*, *N. fischeri* is only rarely identified as a human pathogen (Opal et al. 1986; Summerbell et al. 1992; Lonial et al. 1997; Chim et al. 1998); while *A. clavatus* is probably an important allergen and the causative agent of extrinsic allergic alveolitis known as malt worker's lung (Blyth et al. 1977). *A. clavatus* also produces a number of mycotoxins and has been associated with neurotoxicosis in sheep and cattle fed infected grain worldwide (e.g. (Kellerman et al. 1976)). Our phenotypic characterization (Table S1 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec) has shown that both *A. fumigatus* and *N. fischeri* can grow at 42°C, which indicates that *A. fumigatus* may possess other genetic determinants besides thermotolerance that allow it to establish a successful *in vivo* infection.

As determined by multilocus sequence comparison, most *A. fumigatus* isolates, including Af293 and A1163, lie within the main *A. fumigatus* clade and persist as a single, global phylogenetic population, presumably due to its small spore size (Pringle et al. 2005). Natural *A. fumigatus* isolates were described previously as having low genetic diversity in comparison to *N. fischeri* isolates (Rydholm et al. 2006). However recent studies identified a number of strain-specific (Nierman et al. 2005) and polymorphic (Balajee et al. 2007) genes. To further explore the extent of genetic variation within the *A. fumigatus* species, we included in this analysis the genome sequence of a second strain, A1163, made available through Merck & Co., Inc., Whitehouse Station, NJ. Our preliminary analysis has shown that Af293 and A1163 isolates vary greatly in their resistance to antifungals (Table S2 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec).

7.4 Methods

7.4.1 Fungal Isolates

A. fumigatus Af293 (FGSC A1100) was isolated from patients with invasive aspergillosis (Pain et al. 2004). *A. fumigatus* A1163 (FGSC A1163) is a derivative of *A. fumigatus* CEA17 converted to *pyrG+* via the ectopic insertion of the *A. niger pyrG* gene (d'Enfert 1996; Weidner et al. 1998). CEA17 is a uracil auxotroph of *A. fumigatus* clinical isolate CEA10 (CBS144.89). The type strains of *A. clavatus* (NRRL 1) and *N. fischeri* (NRRL 181) were used for sequencing and phenotypic characterization.

7.4.2 Accession Numbers

The genome sequences of *A. clavatus*, *N. fischeri* and *A. fumigatus* A1163 were deposited to the GenBank under the following accession numbers: **AAKD00000000**, **AAKE00000000** and **ABDB00000000**, respectively.

7.4.3 Whole Genome Sequencing

A1163, *A. clavatus* and *N. fischeri* were sequenced using the whole genome shotgun method as previously described (Fleischmann et al. 1995). Random shotgun libraries of 2-3 Kb, 8-12 Kb and 50 Kb were constructed from genomic DNA from each strain, and DNA template was prepared for high-throughput sequencing using Big Dye Terminator chemistry (Applied Biosystems). Sequence data was assembled using Celera Assembler. For *A. fumigatus* A1163, scaffolds were compared to those of the first sequenced isolate, Af293 (Nierman et al. 2005).

7.4.4 Sequence identity at the nucleotide level

A1163 assemblies larger than 5 Kb were aligned to the Af293 chromosomes using the MUMmer package (http://mummer.sourceforge.net/) (Delcher et al. 1999).

Alignments longer than 100 Kb were used to determine average sequence identity to avoid highly repetitive and duplicated regions. The same approach was used to estimate sequence identity between *A. flavus* and *A. oryzae* and between the two sequenced *A. niger* strains.

7.4.5 Gene structure annotation

The JCVI eukaryotic annotation pipeline was applied to the A1163, *A. clavatus* and *N. fischeri* assemblies (supercontigs) larger than 2 Kb as described earlier (Nierman et al. 2005). We used PASA (Haas et al. 2003) and EvidenceModeler (Haas et al. 2008) to generate consensus gene models based on predictions from several types of genefinders including GlimmerHMM, Genezilla, SNAP, Genewise and Twinscan. Putative pseudogenes, small species-specific genes (less than 50 amino acids), and gene models overlapping with transposable elements (TE) shown in Table S16 were excluded from the final gene lists.

7.4.6 Repetitive elements

Identification of repeat elements was performed using RepeatMasker (http://www.repeatmasker.org/), RepeatScout (http://repeatscout.bioprojects.org/), and Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.html). Putative TEs (Table S16 http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec) were identified by Transposon-PSI (http://transposonpsi.sourceforge.net), a program that performs tBLASTn searches using a set of position specific scoring matrices (PSSMs) specific for different TE families. TE and repeat densities were calculated as the percentage of nucleotide bases in the regions of interest (i.e., syntenic or non-syntenic blocks) that overlap with a feature of the appropriate type (repeat or TE).

7.4.7 *A. fumigatus* annotation improvements

We leveraged the comparative genomic data to significantly improve annotation quality of the Af293 genome, which was previously annotated with relatively little

supporting evidence (Nierman et al. 2005). The refinement of initial annotation was performed using the Sybil software package (http://sybil.sourceforge.net/), which allows for rapid identification of discrepancies in gene structure among orthologs. The comparison with orthologous *N. fischeri* and *A. clavatus* genes resulted in significant changes to the Af293 gene catalogue. Over 1100 gene models were updated and 130 new genes were identified. Initial *A. fumigatus* A1163 gene models were also improved using the PASA pipeline, initially developed to align expressed sequence tag (EST) data onto genomic sequences (Haas et al. 2003). The pipeline was adapted to automatically update A1163 gene models by aligning them against Af293 coding sequences (CDSs).

7.4.8 Functional annotation

We have performed transitive functional annotation from Af293 proteins to their A1163, *N. fischeri* and *A. clavatus* orthologs. Previously GO terms (Ashburner et al. 2000) were assigned to Af293 proteins based on sequence similarity to PFAM domains or experimentally characterized *S. cerevisiae* proteins (Nierman et al. 2005). Secondary metabolism gene clusters were identified using Secondary Metabolism Unique Region Finder (SMURF) available at http://www.jcvi.org/smurf (Chapter 3). The complete list of gene clusters can be downloaded at ftp://ftp.jcvi.org/pub/software/smurf/. Gene Ontology (GO) terms (Ashburner et al. 2000) were assigned as described in (Nierman et al. 2005).

7.4.9 Ortholog identification

After extensive computational and manual refinement, the improved protein datasets were used to generate the final set of orthologs. Orthologous groups in *Aspergillus* genomes were identified using a reciprocal-best-BLAST-hit (RBH) approach with a cut-off of 1e-05. In addition to the A1163, *A. clavatus* and *N. fischeri* genomes, the previously sequenced genomes of Af293 (Nierman et al. 2005), *A. terreus* NIH2624 (http://www.broad.mit.edu), *A. oryzae* RIB40 (Machida et al. 2005), *A. nidulans* FGSC A4 (Galagan et al. 2005) and *A. niger* CBS 513.55 (Pel et al. 2007) were included in the comparative analysis. The

results of this analysis, as well as synteny visualisation and comparative analysis tools can be also found in the *Aspergillus* Comparative database at http://www.tigr.org/sybil/asp. Orthologous, unique and divergent genes in Af293 were identified based on alignments of Af293 CDSs against A1163 assemblies using gmap as implemented in PASA (Haas et al. 2003) using default parameters.

7.4.10 Synteny analysis

Syntenic blocks for each pair of genomes (Af293 vs. *A. clavatus* and Af293 vs. *N. fischeri*) were defined as areas containing a minimum of five matching (orthologous) genes with a maximum of 20 adjacent non-matching genes (having no orthologs) in the reference and target genomes. Since most syntenic regions slightly overlapped, the original blocks were merged to calculate repeat and TE density. Af293 non-syntenic blocks were defined as areas excluded from the syntenic blocks and containing at least ten Af239 non-matching genes.

7.4.11 Statistical analysis

Genes in four lineage-specificity groups were analyzed by the EASE module (Hosack et al. 2003) in MEV within TM4 (http://TM4.org) (Saeed et al. 2006) to identify overrepresented Gene Ontology (GO) terms, Pfam domains and Chromosomal Regions (telomere-proximal and central). Only categories with Fisher's exact test probabilities above with $P > 0.05$ from the EASE analyses were reported for each gene set.

7.4.12 Selective constraints

Selective constraints were estimated for sets of orthologous genes from the Af293, A1163, *A. clavatus*, *N. fischeri* and *A. terreus* genomes. The rate of substitution in synonymous (d_S) and in non-synonymous (d_N) sites, and their ratio (d_N/d_S) was calculated using the PAML package (Yang 1997b). If a gene is very well conserved, $d_N/d_S < 0.1$; if a gene is under weak purifying selection, $0.1 < d_N/d_S < 1$; if a gene is evolving neutrally (e.g. pseudogenes), $d_N/d_S \sim=1$; and if a gene is

evolving under diversifying selection, $d_N/d_S > 1$. The results are reported only for orthologous genes sets having unsaturated d_S values, the same number of exons, and sequence alignment coverage > 95%. For each gene, the average d_N/d_S ratio for five pairwise species comparisons was calculated.

7.4.13 Phylogenetic analyses

We assembled a local database of protein sequences from the 28 publicly available fungal genome projects (Table S17 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec). All phylogenetic analyses in this paper were carried out on protein sequences. The *A. niger* ATCC 1015, *Nectria haematococca*, *Phanerochaete chrysosporium* and *Trichoderma reesei* genomes projects was completed 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 (Contract No. W-7405-Eng-48), Lawrence Berkeley National Laboratory (contract No. DE-AC03-76SF00098) and Los Alamos National Laboratory (contract No. W-7405-ENG-36).

To produce a reference tree of species phylogeny we used the protein sequences of 90 likely orthologs from *A. niger*, *A. nidulans*, *A. terreus*, *A. oryzae*, *A. clavatus*, *N. fischeri*, *A. fumigatus* and *Fusarium graminearum* (teleomorph of *Gibberella zeae*) as an outgroup. To minimize the effect of incorrect or incongruent gene models, these proteins were chosen on the basis of having identical numbers of introns in each species and similar lengths. Sequences were aligned using MUSCLE (Edgar 2004) and columns of low conservation were removed manually. Maximum-likelihood trees were constructed using the PHYLIP package, applying the JTT substitution model with a gamma distribution (alpha = 0.5) of rates over four categories of variable sites.

Phylogenetic analyses of individual Af293, A1163, and *N. fischeri* proteins were carried out on sets of homologs identified in BLASTP searches against our fungal

database. The top 20 hits with $E < 10^{-4}$ were retained for analysis. Sequences were aligned using ClustalW (Thompson et al. 1994). Poorly aligned regions were removed using Gblocks (Castresana 2000). Finally, a maximum likelihood tree was drawn using PHYML (Guindon and Gascuel 2003).

7.4.14 Southern blot analysis

To detect polymorphisms in the *rosA* (AFUA_6G07010) gene, several hybridizations were performed using *rosA* gene as the probe and genomic DNA cleaved with *EcoR*I, *ClaI*, *BamH*I or *EcoR*V. For comparison, an invariable gene for all species (*apg5*; AFUA_6G07040) was used as the hybridization probe on genomic DNA digested with *Hpa*I.

7.4.15 Colony radial growth rate measurement

Colony radial growth rate measurements were performed as described (Robson et al. 1995). For each isolate, four (90 mm diameter) Petri dishes containing 25 ml agar medium were inoculated centrally with 2.5 microlitres of 1×10^6 spores/ml suspension in PBS/Tween 80. Plates were then incubated at temperatures ranging from 25°C to 50°C and colony edges were marked using a plate microscope. Colonies were marked twice daily for 4-5 days. For each colony, two diameters perpendicular to each other were measured.

Eight replicates were measured for each isolate. The results reported here are the mean of two experiments. At least five time points during the log phase were used to calculate growth rate. The radius of the colonies was plotted against time using least-square regression analysis, and the slope of the regression line, which represents the growth rate, was calculated. Each replicate was analysed separately and the mean of the growth rate was then calculated.

7.5 Results and discussion

7.5.1 *A. fumigatus* Af293 vs. *A. fumigatus* A1163

The genome of *A. fumigatus* strain A1163 was sequenced by the whole genome random sequencing method (Fleischmann et al. 1995). Its genome (29.2 Mb) is 1.4% larger than the genome of the first sequenced strain Af293 (28.8 Mb) (Table 7.1Table 7.1Table 7.1). About 98% of each genome can be aligned with high

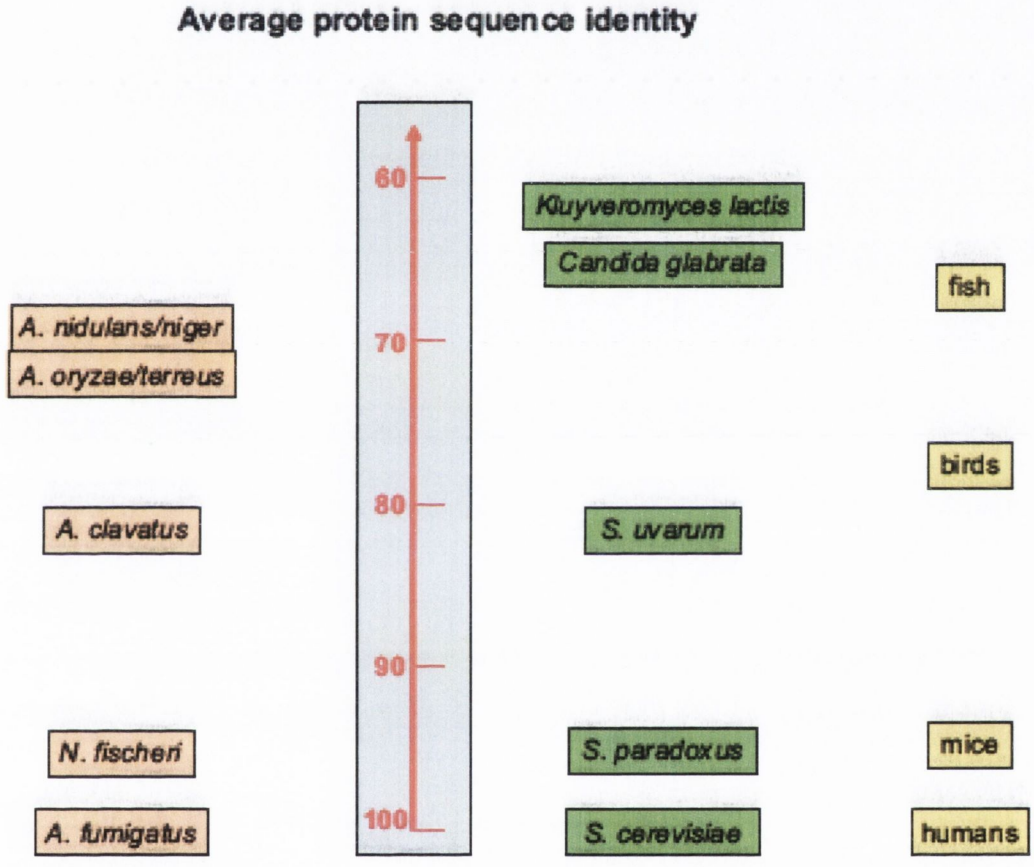


Figure 7.1 Molecular Divergence in Molds and Yeasts. *A. fumigatus* proteins are compared to their orthologs in *N. fischeri*, *A. clavatus*, *A. terreus*, *A. oryzae*, *A. nidulans*, and *A. niger* (mean values: 95%, 84%, 71%, 71%, 68%, and 69%, respectively). Saccharomyces paradoxus, *Saccharomyces uvarum*, *Candida glabrata*, and *Kluyveromyces lactis* are compared to *Saccharomyces cerevisiae* (adapted from [74,75]). Mean values for these species are 90%, 82%, 64%, and 60%, respectively. Median percent identity between pairs of orthologs from *A. fumigatus* and each successive genome in the tree is shown. Relative divergence of humans, mice, birds and fish are shown for reference.

confidence. Alignment of the A1163 genome against the eight Af293 chromosomes has revealed 17 large syntenic blocks, which correspond roughly to the 16 Af293 chromosomal arms (Figure 7.3). The syntenic blocks were defined as regions containing at least five syntenic orthologs separated by no more than 20 genes without orthologs.

Most translocation events involving *A. fumigatus* chromosomes appear to have taken place within 300 Kb from the telomeres. The largest exchange involved a ~500 Kb segment between Af293 chromosomes 1 and 6 and A1163, which contain regions aligning with A1163 assembly 1 (syntenic blocks 1.1 and 1.2 in Figure 7.3). This appears to be a recent event that happened in A293. In addition, Af293 chromosome 1 harbours a 400 Kb subtelomeric region that does not align well with A1163 assemblies. There is evidence of gene conversion between distal subtelomeric sequences encoding RecQ family helicases in *A. fumigatus* chromosomes 2, 4, and 7.

Consistent with previous reports (Rydholm et al. 2006), the identity over the shared regions is very high (99.8% at the nucleotide level). This is higher than 99.3% and 99.5% identity between the two sequenced *A. niger* isolates (ATCC 1015 and CBS 513.88) (Pel et al. 2007) and between *A. oryzae* (Machida et al. 2005) and *A. flavus* (Payne et al. 2006), respectively. Unique regions represent 1.2% and 2.3% (and harbour 143 and 218 genes) in the Af293 and A1163 genomes, respectively. More than half of the Af293-specific genes are also absent in *A. fumigatus* isolates Af294 and Af71, according to the array-based comparative genome hybridization (aCGH) data (Nierman et al. 2005). The vast majority of Af293- and A1163-specific genes are clustered together in blocks ranging in size from 10 to 400 Kb, which seem to be the most variable segment of the species genome. A manual examination of these isolate-specific islands revealed that they contain numerous pseudogenes and repeat elements. One of the regions contains a putative secondary metabolism cluster (AFUA_3G02530-AFUA3G02670).

The origin of 20% of Af293-specific genes can be attributed to two segmental duplication events. One of the duplicated regions (AFUA_1G16010-AFUA_1G16170) contains an arsenic detoxification cluster. The other (AFUA_1G00420-AFUA_1G00580) contains genes that may be involved in metabolism of betaine, which is often synthesized under osmotic and heavy metal stress. Interestingly the duplicated regions are also absent in Af294 and Af71 isolates, which suggests that the duplication event took place very recently.

Segmental duplication events are thought to contribute to rapid adaptation of the species by increasing their expression. Since Af293 is a clinical isolate it is possible that these chromosomal aberrations were created due to selective pressures in the host.

7.5.2 Highly Variable Loci in A. fumigatus

Although most Af293 proteins are 100% identical to their A1163 orthologs, we have identified 41 orthologous pairs that share only 37% to 95% identity. To find out if these genes are also divergent in other *A. fumigatus* isolates, we identified Af293 genes that do not hybridize with DNA extracted from the Af294 and Af71 strains in aCGH experiments (Nierman et al. 2005). The comparison revealed that 27 out of 41 genes were possibly polymorphic (marked as absent or divergent) with respect to at least one other isolate (Table S3 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec). Further analysis of three polymorphic loci in other *A. fumigatus* isolates has demonstrated that each of them harbours two or three alleles (Table S4 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec). A PCR survey followed by Southern blot analysis and partial DNA sequencing has shown the presence of at least two alleles at each locus containing nearly identical sequences within each group of alleles (data not shown).

Table 7.1 Genome Statistics

Sequenced organisms	***Af293***	***A1163***	***N. fischeri***	***A. clavatus***
Length (Mp)	28.810	29.205	32.552	27.859
Assemblies > 100 Kb	18	11	13	16
GC content	50%	49%	49%	49%
No. of genes	9631	9906	10407	9125
No. of LS genes	818		1408	1151
Mean gene length (Bp)	1478	1455	1466	1483
% Genes with introns	79%	80%	80%	81%
% Coding	49%	49%	47%	49%

In filamentous fungi, this high level of variability has been previously associated with heterokaryon incompatibility (*het*) genes involved in a programmed cell death (PCD) pathway triggered by hyphal fusion between two genetically incompatible individuals (Glass and Dementhon 2006; Paoletti et al. 2007). So far several *het* loci have been described in *A. nidulans* (Anwar et al. 1993), although none have been characterized at the molecular level. Incidentally, our results are consistent with previously identified vegetative incompatibility groups suggesting that some of these polymorphic genes may function in heterokaryon incompatibility in *A. fumigatus*. Thus, four clinical isolates from the same multi-member incompatibility group (WSA-270, WSA-1195, WSA-449, and WSA-172) contained the same alleles of the polymorphic genes (Table S4 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec).

Furthermore, at least five putative *A. fumigatus het* genes exhibit a pattern of trans-species (or trans-specific) polymorphism (Table S5 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec), which has been previously associated with somatic and sexual incompatibility in fungi, self-incompatibility in plants, and the major histocompatibility complex (MHC) in vertebrates. These genes are more similar to their orthologs from other *Aspergillus* species than to those from A1163. We chose one putative *het* gene, *rosA* (AFUA_1G15910), and its close relative, *nosA* (AFUA_4G09710), whose orthologs encode two Zn2C6 transcriptional regulators of sexual development in *A. nidulans* (Vienken et al.

2005; Vienken and Fischer 2006) for phylogenetic analysis (Figure 7.4). Unexpectedly, Af293 RosA clusters with its *A. clavatus* ortholog, while A1163 RosA clusters with *N. fischeri*. This is in contrast with the NosA tree, which perfectly mirrors the species tree (Figure 7.2), suggesting that these allelic classes may transcend species boundaries in the aspergilli.

This is the first study that shows the diversity of *het* genes in aspergilli at the molecular level as well as patterns of trans-species polymorphism. These putative *het* genes are distinct from those identified in *Neurospora crassa* or *Podospora anserina* [24,25], although many of them share the same domains such as the NACHT and NB-ARC domains of the STAND superfamily (Leipe et al. 2004). Coincidentally four of the *A. fumigatus* variable genes encoding STAND domain proteins have previously been predicted to function in heterokaryon incompatibility (Fedorova et al. 2005). The discovery of putative *het* loci in the aspergilli may facilitate identification of downstream components of fungal PCD pathways or other drug targets. These loci may be also used as a basis for classification of natural and clinical isolates into different compatibility groups.

7.5.3 *A. fumigatus* vs. *N. fischeri* vs. *A. clavatus*

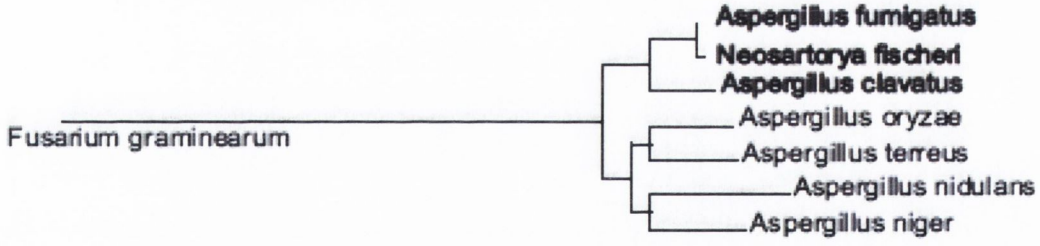


Figure 7.2 Three Closely Related Aspergilli. The three most closely related aspergilli, which constitute the Affc-core group (A. fumigatus, N. fischeri, and A. clavatus), are in bold black. The maximum-likelihood tree was constructed from an alignment of 90 proteins chosen on the basis of similar lengths and identical number of intron/exon structures in order to minimize the number of inconvenient or incongruent gene models (see Methods7.4).

The genomes of *N. fischeri* and *A. clavatus* were sequenced by the whole genome sequencing method (Fleischmann et al. 1995). The *N. fischeri* genome (32.6 Mb) is 10-15% larger than the *A. clavatus* and *A. fumigatus* genomes (Table 7.1).

Table 7.2 Syntenic and Afum-specific Chromosomal Blocks in Af293

Af293 blocks	Syntenic to A1163	Syntenic to *N.fischeri*	Syntenic to *A.clavatus*	Afum specific
No. of original blocks	29	24	62	13
No. of merged blocks	17	20	55	13
Merged blocks length	28.4 Mb	27.6 Mb	26.0 Mb	1.7 Mb
% Coding	50%	51%	52%	31%
Repeat[a] density	0.51%	0.50%	0.47%	1.83%
TE[b] density	1.07%	0.96%	0.80%	4.17%

Note- Syntenic blocks for each pair of genomes were defined as areas containing a minimum of five orthologous genes in the Af293 and target genomes with a maximum of 20 adjacent non-matching genes. Afum-specific blocks were defined as Af293 areas containing at least ten Afum-specific genes and separated by no more than 5 other genes. Since most syntenic regions slightly overlap, the original blocks were merged to calculate repeat and TE density. Abbreviations: [a]repeat elements; [b]transposable elements. Repeat and TE densities were estimated as described in Methods.

There are 10,407 protein-coding genes and a large number of transposable elements, which may have contributed to its genome size expansion. The *A. clavatus* genome (27.9 Mb) is the smallest seen to date among the sequenced aspergilli (Table 7.1). There are currently 9,125 predicted protein-coding genes. This is consistent with past comparative studies that identified notable (up to 30%) genome size differences between distantly related aspergilli (Galagan et al. 2005; Machida et al. 2005; Nierman et al. 2005).

Despite this significant genome size variability, gene-level comparisons confirmed phylogenetic proximity of *A. fumigatus*, *N. fischeri* and *A. clavatus* (Figure 7.1 and Figure 7.2). The three genomes also appear to be largely syntenic. Alignment of the *N. fischeri* and *A. clavatus* genomes against the eight Af293 chromosomes has revealed 20 and 55 syntenic blocks, respectively (Table 7.2). There is only one large-scale reciprocal translocation between chromosomes

2 and 5 in *N. fischeri* (blocks 8927.1, 8927.2, 9292.1 and 9292.2, in Figure 7.3). The *A. clavatus* supercontigs align with *A. fumigatus* chromosomes 2 and 5, suggesting that this was the ancestral topology.

7.5.4 Core and Lineage-Specific Genes

7.5.4.1 Features of Core and Lineage-Specific Genes

Comparative genomic analysis has showed that the three *Aspergillus* genomes contain a large number of species-specific genes, which is consistent with previous comparative studies (Nierman et al. 2005). We have identified 7514 orthologous core and 818, 1402 and 1151 species-specific genes in the Af293, *N. fischeri* and *A. clavatus* genomes, respectively (Figure 7.5). Numbers of core- and species-specific genes, however, depend on selection of genomes from which they were derived. Thus, adding new genomes to this comparison resulted in fewer core and specific genes as shown for Af293 in Table S6. The availability of additional sequenced *Aspergillus* genomes allowed us to explore these patterns in a more systematic manner by comparing *A. fumigatus* Af293 genes with different lineage specificity (i.e. number of orthologs in other species).

To this end, we have selected four sets of genes based on the presence of orthologs in the six other sequenced aspergilli: *N. fischeri*, *A. clavatus, A. terreus* (CH476594), *A. oryzae* (Machida et al. 2005) *A. nidulans* (Galagan et al. 2005) and *A. niger* CBS 513.88 (Pel et al. 2007) (Table S6 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec; Figure 7.5). Genes with orthologs in the three most closely related aspergilli (*A. fumigatus, N. fischeri* and *A. clavatus*) constitute the Affc-core group. The genes in the Affc-core can be further divided into two groups, the *Aspergillus*-core (Asp-core) with orthologs in all six other aspergilli and the Affc-specific group, which is comprised of the remaining Affc-core genes. Finally, the *A. fumigatus*-specific (Afum-specific) group contains Af293 genes that have orthologs in neither *N. fischeri* nor *A. clavatus*.

One of the most striking observations to arise from this comparison was the marked differences in size and number of exons among genes from different lineage-specificity groups (Table 7.3). For example, Asp-core genes on average are almost twice as large as Afum-specific genes. The latter have on average only 1.35 introns and almost 31% lack introns completely. In contrast, Asp-core genes contain on average 2.16 introns, only 16% of them without introns. Consistent with previous reports of increased evolutionary rates in LS genes (e.g. (Cai et al. 2006)), Affc- and Afum-specific genes in *A. fumigatus* exhibit low sequence identity to their orthologs from more distantly related fungi (Table 7.3).

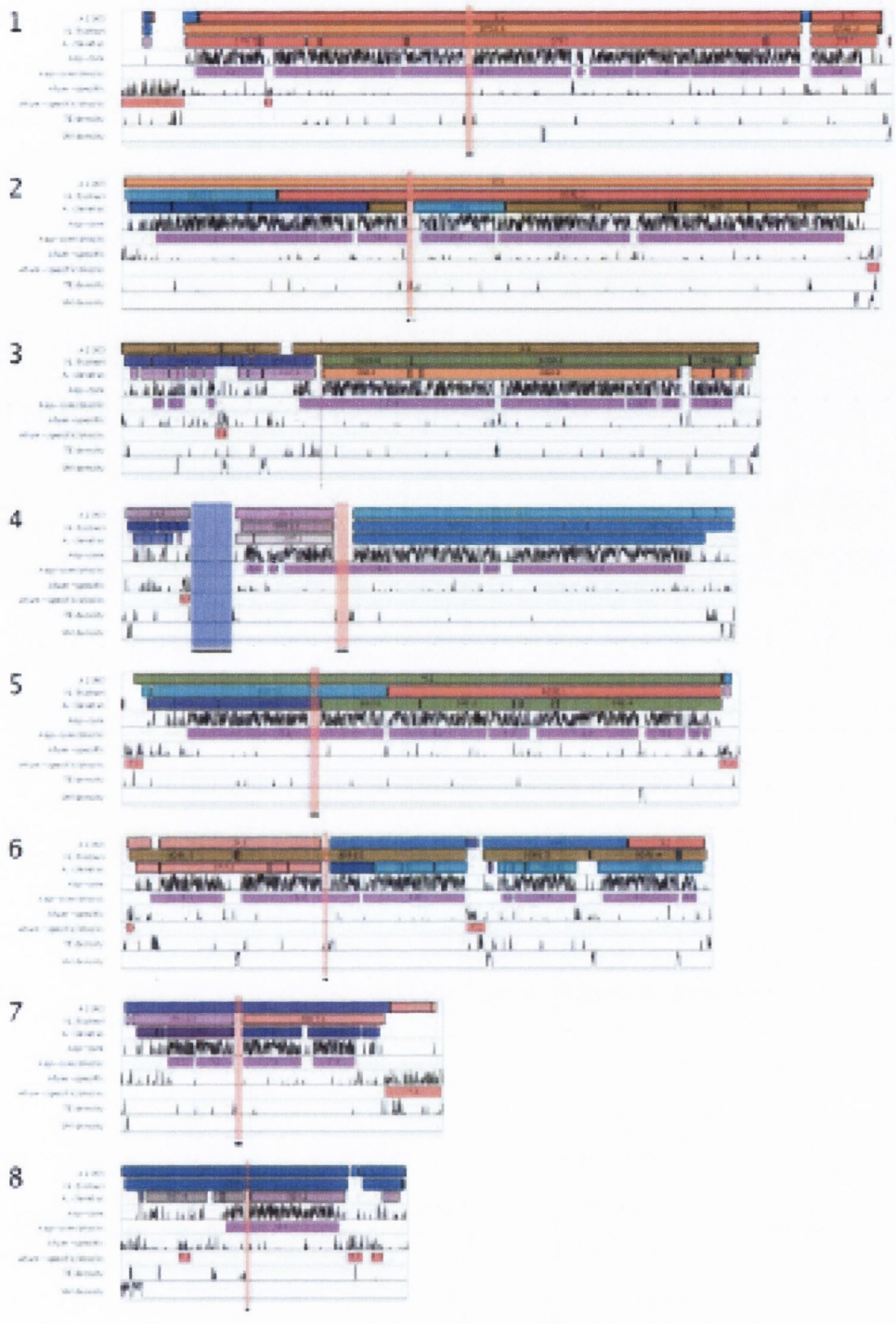


Figure 7.3 Alignment of the A1163, N. fischeri, and A. clavatus Assemblies against the Eight Af293 Chromosomes. The first three tracks from the top for each reference chromosome show syntenic blocks (horizontal bars) identified in the target genomes, A. fumigatus A1163, N. fischeri, and A. clavatus. Each assembly from the target genomes is represented by a single color. Syntenic blocks are numbered based on the target genome assembly ID and the position of the block in the target genome assembly. Tracks 4 and 5 show Asp-core gene density and blocks (horizontal bars), respectively, in the Af293 genome. Tracks 6 and 7 show Afum-specific gene density and blocks (horizontal bars), respectively. Tracks 8 and 9 show the density of clustered secondary metabolite biosynthesis genes and transposable elements, respectively, found in Af293. Pink vertical bars represent putative centromeres, the purple vertical bar in chromosome 4 represents a region of ribosomal DNA, and horizontal black bars beneath each chromosome designate sequencing gaps.

These vast differences in gene features between core and specific genes are more likely to be explained by relaxed selective constraints (as discussed below) than by poor annotation quality of LS genes (due to misannotated gene models, gene fragments or random ORFs). We made significant improvements to Af294 gene models by leveraging the comparative genomic data (see Methods 7.4). In

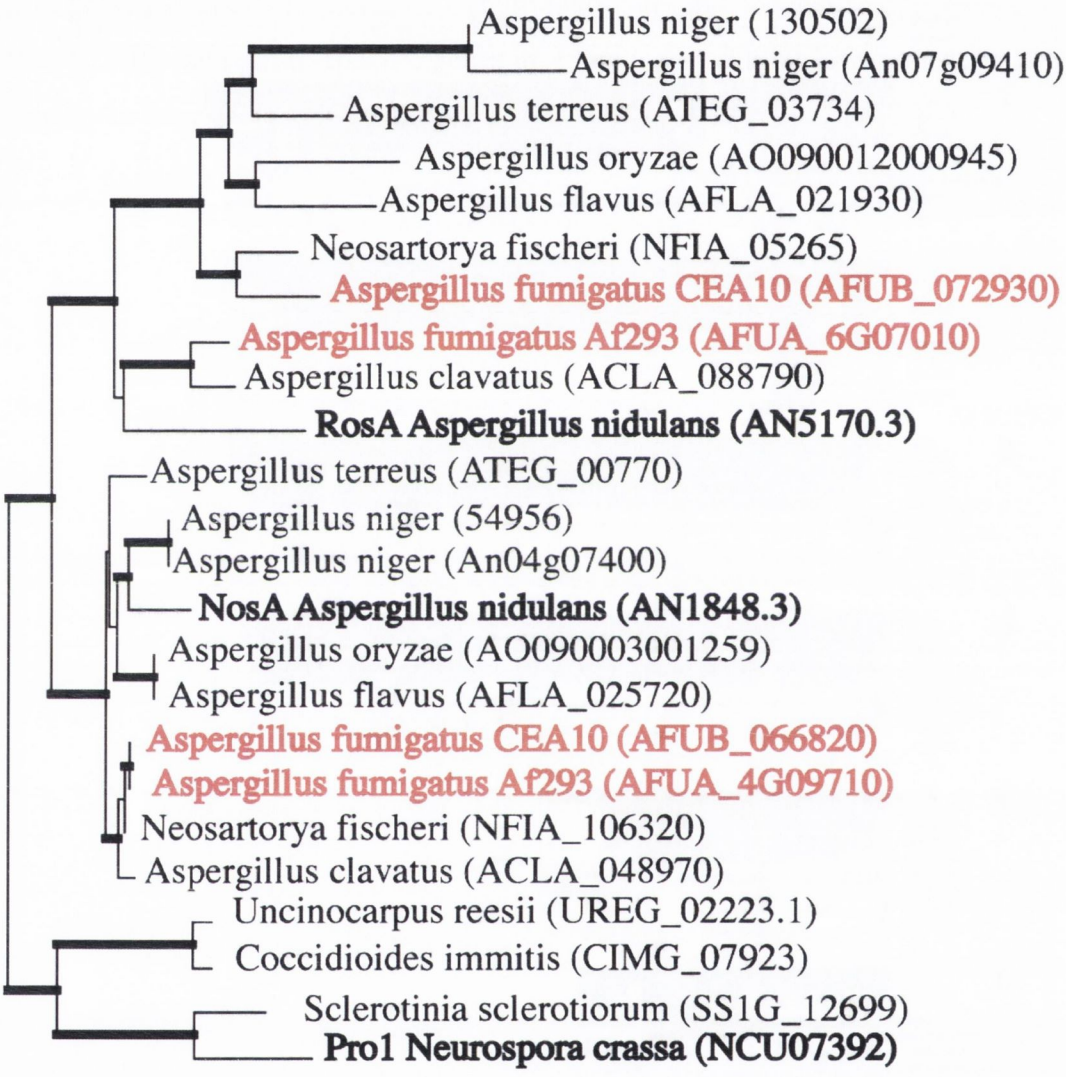


Figure 7.4 The Af293 RosA and NosA Proteins. Shown in bold red are RosA, NosA and Pro1 proteins that have been experimentally characterized are shown in bold black. Branches with a bootstrap of 75% or more are indicated in bold black. The trees are maximum-likelihood trees (see Methods 7.4).

addition, all Affc-specific genes have orthologs in *N. fischeri* and *A. clavatus* and 43% of them are differentially expressed in various expression studies, which is similar to the *A. fumigatus* genome average (Table 7.3). On the other hand, many Afum-specific genes may be non-functional, since only 32% of them are differentially expressed in microarray studies (vs. the 43% genome average) and

only 60% of them show sequence similarity to other fungal proteins (Table S7 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec; Figure 7.6). Nonetheless, at least 20% of Afum-specific genes are supported by combined evidence (homology and expression data) and therefore are likely to be functional. Nonetheless, even these genes are still smaller in size than average Affc- and Asp-core genes.

7.5.4.2 Biological Roles and Chromosomal Location of LS Genes

Analysis of Gene Ontology (GO) terms (Ashburner et al. 2000) associated with core and lineage-specific groups has demonstrated that certain biological

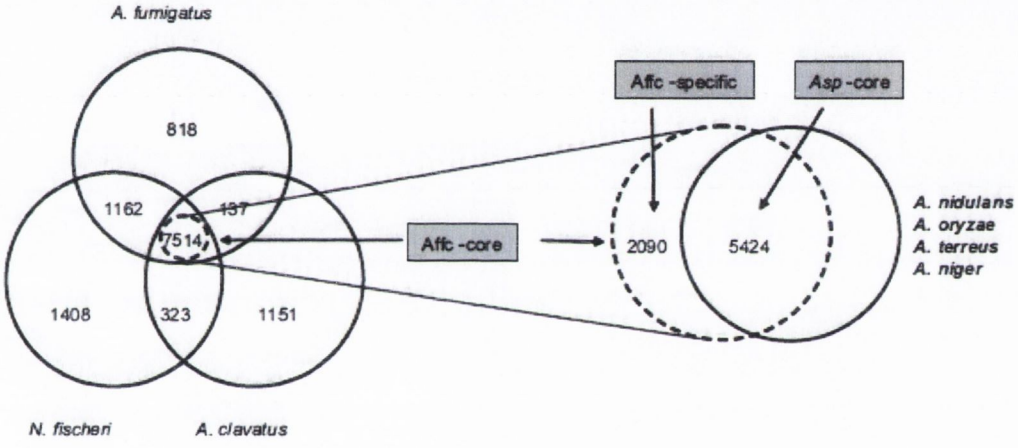


Figure 7.5 Proteins with Orthologs in the Three Most Closely Related Aspergilli (*A. fumigatus*, *N. fischeri* and *A. clavatus*). These proteins constitute the Affc-core group, and proteins with no orthologs in *N. fischeri* and *A. clavatus* constitute the A. fumigatus-specific group (Afum). The proteins in the Affc-core can be further divided into two groups, Aspergillus-core (Asp-core), which has orthologs in all of the other aspergilli, and the Affc-specific group, which is comprised of the rest of the Affc-core.

functions are unequally distributed among these groups (Table S8 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec). The Afum-specific group is enriched for genes involved in carbohydrate transport and catabolism, secondary metabolite biosynthesis, and detoxification. In contrast, the invariable Asp-core genome encodes many functions associated with information processing and other cellular processes that contribute to the organism's fitness in most environments. Thus, a significant number of Asp-core genes (15%) are orthologous to yeast essential

genes, which represents a two-fold enrichment in comparison to the rest of the proteome.

Although most Af293 genes involved in carbohydrate transport and catabolism are found in the Asp-core group, only 10% of secondary metabolism genes have orthologs in all sequenced aspergilli including siderophore, pigment and Pes1-related clusters. These three conserved clusters are also found in *Penicillium*

Table 7.3 Comparison of Four Af293 Gene Sets with Different Lineage Specificity

Lineage specificity group	Asp-core	Affc-core	Affc-specific	Afum-specific
No of genes	5424	7514	2090	818
No of orthologs in 6 aspergilli	6	2-6	2-5	0-1
Mean gene length	1722	1579	1209	802
Mean No. of introns	2.16	2.02	1.66	1.35
%Genes without introns	15.9%	19.4%	28.5%	31.4%
% Affc syntenic	98.3%	96.0%	89.8%	n/a
% Telomere-proximal	5.6%	9.1%	38.0%	36.5%
% Expressed	42.5%	42.7%	43.3%	32.4%
% Orthologs in *A.clavatus*	100%	100%	100%	n/a
% Orthologs in *N.crassa*	81.5%	70.7%	42.6%	4.5%
% Orthologs in *S.cerevisiae*	49.9%	41.5%	19.9%	1.2%
% Identity to *A.clavatus* orthologs	81.3%	78.6%	71.4%	n/a
% Identity to *N.crassa* orthologs	52.3%	51.6%	47.9%	43.3%
% Identity to *S.cerevisiae* orthologs	43.1%	42.7%	40.4%	38.0%

Note- The numbers of Af293 genes in different categories are shown for *Aspergillus*-core (Asp-core), Affc-core, Affc-specific, and *A. fumigatus*-specific (Afum-specific) groups (see main text for definitions). Telomere-proximal genes are defined as genes located within 300 Kb from the chromosome end. Affc syntenic genes are defined as Af293 genes syntenic with respect to *N. fischeri* and *A. clavatus* (see the legend to Table 7.2). The 'expressed' genes are defined as Af293 genes that showed differential expression in at least one microarray study (W. Nierman, unpublished).

species and some more distantly related fungi. Similarly, only 30% of secondary metabolism Af293 genes are shared by *N. fischeri* and *A. clavatus*. The three species also vary considerably in the numbers of enzymes that control the first step in secondary metabolite biosynthesis such as nonribosomal peptide synthases (NRPS), polyketide synthases (PKS), and dimethylallyltryptophan synthases (DMATS) (Table S9 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec). Interestingly, *N. fischeri* genome contains 46 enzymes, which is 35% more than *A. clavatus* (35) and *A. fumigatus* (34) genomes.

Likewise, PFAM domains overrepresented among Affc- and Afum-specific genes have been shown to function in efflux or detoxification, secondary metabolite biosynthesis, resistance to antifungals, and other accessory metabolic pathways. They include MSF and ABC transporters, various oxidoreductases, cytochrome P450, glycosyl and alpha/beta fold hydrolases, polyketide synthases, glutathione transferases and methyltransferases (Table S10 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec). On the other hand, core genes often contain AAA-superfamily ATPase, helicase, WD40, and SH3 domains associated with such important functions as cell organization and macromolecule biosynthesis.

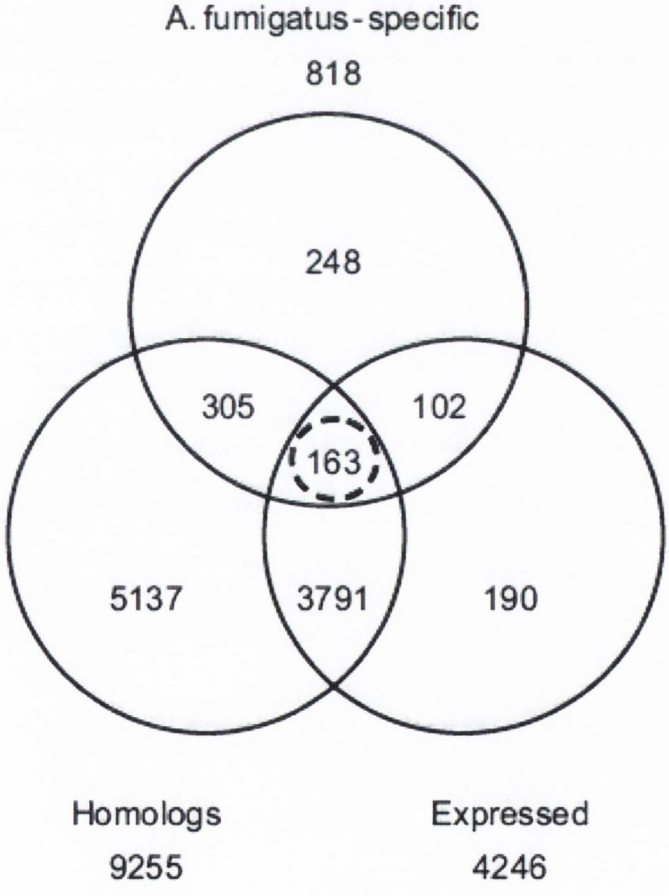


Figure 7.6 *A. fumigatus*-Species Specific Genes Supported by Homology and Expression Data. Genes with no orthologs in *N. fischeri* and *A. clavatus* constitute the *A. fumigatus*-specific group (Afum). Genes that have homologs in other fungal genomes constitute the Homology group. Genes differentially expressed in microarray studies represent the Expressed group.

7.5.4.3 Lineage Specific Genomic Islands

In addition to difference in size and function, lineage specific genes display a significant subtelomeric bias. As opposed to telomere-distal Asp- and Affc-core genes, Affc- and Afum-specific genes tend to be located within 300 Kb from chromosome ends (P value > 0.01) (Table S11 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec). About 38% of Affc-specific genes are telomere-proximal in comparison to 6% of Asp-core and 9% Affc-core genes (Table 7.3 Table 7.3 Table 7.3 Table 7.3 Table 7.3). Interestingly, 46% of Afum-specific genes with paralogs are telomere-proximal (Table S7 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec), suggesting that they may have been recently duplicated and translocated to these regions. Our findings concur with previous reports of subtelomeric bias in LS genes in *A. fumigatus* (Nierman et al. 2005), *S. cerevisiae* (Zakian 1996) and *Pichia stipitis* (Jeffries et al. 2007). With the exception of one Af293 locus containing four P450 genes, the *Aspergillus* species do not have large variable subtelomeric arrays arising by a series of tandem duplications found in some protozoan parasites (Berriman et al. 2002).

Almost 50% of the Afum-specific genes can be clustered together in 13 blocks containing more than 10 Afum-specific genes separated by no more than 5 genes outside this category (Table 7.2). Together these regions, referred to here as Afum-specific genomic islands, show an even more significant telomeric bias (68% of the clustered genes lay within 300 Kb from telomere ends) with larger blocks found almost exclusively at chromosome ends (Figure 7.3). In addition to non-syntenic genes, species-specific islands harbour a disproportionate number of transposons and other repeat elements in comparison with the syntenic areas of the Af293 genome (Table 7.3). Notably two *A. fumigatus*-specific blocks (2.2 and 3.1) contain gene clusters involved in biosynthesis of mycotoxin fumigaclavine and another unknown secondary metabolite (Perrin et al. 2007b). Similar genomic islands have been described in the rice blast fungus *Magnaporthe oryzae* (Rehmeyer et al. 2006; Thon et al. 2006) and in *A. oryzae* (Machida et al. 2005) suggesting that they may be shared across all filamentous ascomycota fungi.

Unlike variable subtelomeric regions found in other eukaryotes (Pryde et al. 1997; Mefford and Trask 2002), these areas are often quite large (up to 400 kb) and not always located near chromosome ends.

7.5.4.4 Evolutionary Origins of Lineage-Specific Genes

Most Affc- and Afum-specific genes have no orthologs in non-*Aspergillus* fungal species, which suggests that they were created *de novo* in the Affc lineage. To gain insight into the origin of the LS genes in aspergilli, we have performed phylogenetic analysis of two sets of *A. fumigatus*- and *N. fischeri-specific* genes. In Af293 and *N. fischeri*, Set 1 contains 790 and 1230 genes, respectively, that have an *Aspergillus* homolog as the best BLASTp hit; Set 2 contains 28 and 178 genes, respectively, that have a non-*Aspergillus* homolog as the closest relative. There is a significant difference in the numbers of trees including a non-*Aspergillus* species as the closest relative in *N. fischeri* and *A. fumigatus* (P value = 2.6e-08). This is indicative of major differences in retention and/or uptake of new genetic material in these two species, consistent with differences in their reproductive modes.

The four repetitive scenarios identified by phylogenetic analysis are displayed in Figure 7.7 . In both *A. fumigatus* and *N. fischeri*, most of the Set 1 genes exhibit topologies that do not strictly follow the *Aspergillus* species tree (Figure 7.2), although nested within the *Aspergillus* clade. Similarly, all 28 *A. fumigatus* Set 2 genes are nested within the *Aspergillus* genus. In contrast to the *A. fumigatus* genes, *N. fischeri* Set 2 genes sometimes cluster with a non-*Aspergillus* species with high bootstrap support. As shown in Figure 7.7 B and C, both *N. fischeri* and non-*Aspergillus* species genes can be nested either in this non-*Aspergillus* clade or in the *Aspergillus* clade. At first sight, these repetitive topologies can be interpreted as supportive of a horizontal gene transfer (HGT) from a non-*Aspergillus* species into *N. fischeri* or visa versa. Further analysis, however, reveals that most of the conflicts involve sparsely populated trees, long branch attraction artifacts, and other situations, where phylogenetic methods tend to mislead (e.g. (Rice and Palmer 2006)). The last repetitive scenario includes genes

that are only present in one other distant fungal genome (Figure 7.7 D). The evolutionary origin of genes in this category cannot be resolved at this time.

Our results are consistent with the well established role of gene duplication and divergence as the principal source of new genes (Ohno 1970; Kimura and Ohta 1974; Kurland et al. 2003; Choi and Kim 2007). They are however in conflict with previous studies that attributed the origin of LS genes in the aspergilli to

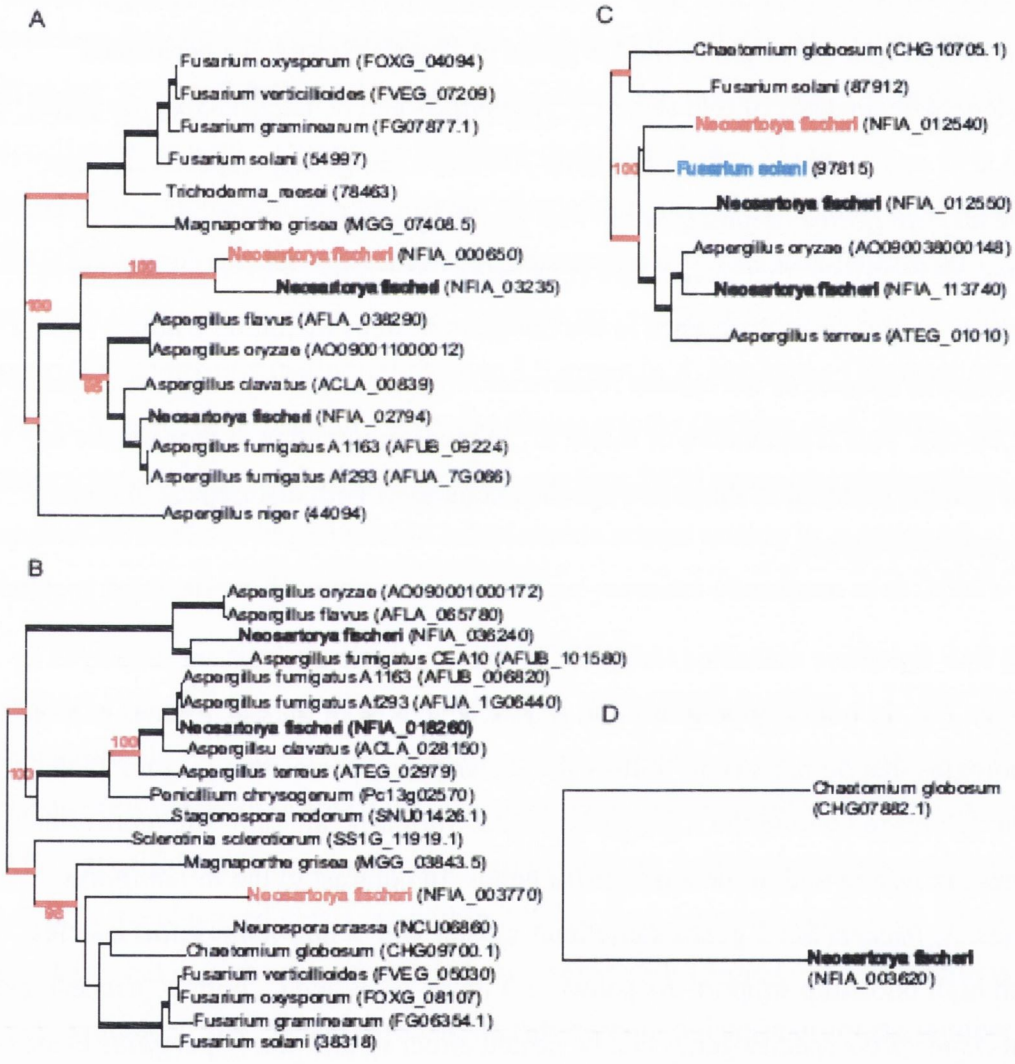


Figure 7.7 Four Common Topologies Detected by Phylogenetic Analysis of *N. fischeri*-Specific Proteins. The *N. fischeri* proteins under consideration are in bold red. The bootstrap supporting the clade containing the *N. fischeri* is also in bold red. Other *N. fischeri* proteins are shown in bold black. Blue species names correspond to the recipient genome when different from *N. fischeri*. Systematic gene names are indicated. Branches with a bootstrap of 75% or more are indicated in bold black. The trees are maximum-likelihood trees (see Methods 7.4). A. Set1 protein evolved by probable duplication, differentiation and differential loss in other *Aspergillus* species (DDL). B. Set 2 protein evolved by probable HGT from Sordaryomycetes into the *N. fischeri* lineage. C. Set 2 protein evolved by probable DDL and a Fusarium solani protein (in blue) evolved by probable HGT from the N. fischeri lineage into Sordaryomycetes. D. Set 2 protein showing similarity to a protein from the Sordaryomyce *Chaetomium globosum*.

gene acquisition through HGT from other fungal species (Galagan et al. 2005; Goffeau 2005; Machida et al. 2005). This assumption was based on circumstantial evidence such as mosaic phyletic distribution, phylogenetic anomalies, and differences in gene content among *A. fumigatus*, *A. nidulans* and *A. oryzae.* Besides the absence of readily apparent HGT examples, the fact that LS genes tend to be smaller in size and have fewer exons is difficult to explain by HGT. These gene features are quite consistent across *Aspergillus* species, and it is therefore unclear what could be the donor organism for LS genes.

The DDL scenario does not have this weakness, since these size differences can be a direct consequence of relaxed selective constraints operating on duplicate genes. According to the DDL hypothesis, the initial redundancy in gene function allows duplicate genes to quickly accumulate nonsynonymous mutations and even premature stop codons. Notably, over 20% of all Afum-specific genes can be linked to the two very recent segmental duplications events that occurred in Af293 but not in A1163. Both translocated segments are telomere-distal and contain genes that appear to be pseudogenized indicating that translocated gene copies may have evolved under relaxed selective constraints. Similarly in other species, accelerated evolution has been often associated with subtelomeric areas suggesting that the process is dependent on the local chromatin environment (e.g. (Kellis et al. 2003)).

The prevailing role of duplication in the origin of LS genes in the aspergilli is further underlined by their tendency to cluster in genomic islands. These regions may function as designated "gene dumps" and simultaneously as "gene factories", since some LS genes appear to maintain their functional integrity or at least are differentially expressed in microarray studies as shown above. As shown above, 46% of Afum-specific genes with paralogs are telomere-proximal (Table S7 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec), suggesting that they may have been recently duplicated and translocated to these regions. Evidence for gene duplication and/or transfer to evolutionarily labile regions is found in some protozoan parasites that have large variable subtelomeric arrays arising by a series of tandem duplications (Berriman et al. 2002).

7.5.5 Conservation of Virulence-, Allergy-, and Sex-Associated Genes

Previous studies however have shown a high level of evolutionary conservation and phyletic retention among known *A. fumigatus* virulence-associated genes (Nierman et al. 2005). Our analysis confirmed the low rate of protein evolution among these genes in four *Aspergillus* species (Table S12 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec). Interestingly, four of the virulence-associated genes, *pabaA* (AFUA_6G04820), *fos*-1 (AFUA_6G10240), *pes*1 (AFUA_1G10380) and *pksP* (AFUA_2G17600), reveal evidence of accelerated evolution in the branch leading to the two *A. fumigatus* isolates. This pattern can affect only a few amino acid residues (e.g. PksP) or a significant proportion of the protein (e.g. Pes1).

Such a pattern can be due to either relaxation of selection or selection for rapid diversification (positive selection). In the latter case specific amino acid substitutions may decrease susceptibility to specific environmental challenges and thus enhance *A. fumigatus* virulence. These four genes are involved in oxidative stress or nutrient availability, which is consistent with the positive selection scenario. Indeed, PabaA is involved in biosynthesis of folate, an essential co-factor for DNA synthesis. Since PABA is apparently limited in the mammalian lung, a functional *pabaA* gene is required for virulence (Brown et al. 2000). Fos1, a putative two-component histidine kinase, may play a role in the regulation of cell-wall assembly (Clemons et al. 2002). Finally, PksP and Pes1 are enzymes, which catalyze the first steps in biosynthesis of the spore pigment and an unknown non-ribosomal peptide, have been shown to mediate resistance to oxidative stress in addition to their role in *A. fumigatus* virulence (Langfelder et al. 1998; Reeves et al. 2006). The inclusion of additional taxa in the analyses might clarify the significance of the observed differences.

This overall lack of variability among known virulence-associated factors suggests that yet unknown *A. fumigatus*-specific genes may contribute to its ability to survive in the human host. A recent microarray study demonstrated that the Affc-specific genes are over-represented among genes that are up-expressed in the neutropenic murine lung (Elaine Bignell, submitted for publication). Many of them are found in chromosomal gene clusters associated with macromolecule catabolism and secondary metabolite biosynthesis. Similarly, clustered lineage-specific genes simultaneously induced in infected tissue have been observed in the ubiquitous maize pathogen *Ustilago maydis* (Kamper et al. 2006) and some other species (for a recent review see (Merrick and Duraisingh 2006)). Alternatively *A. fumigatus* virulence may be a combinatorial process, dependent on a pool of genes, which interact in various combinations in different genetic backgrounds as suggested previously (Nierman et al. 2005). Similar 'ready-made' virulence features have been described in other environmental pathogens such as *Pseudomonas aeruginosa* (Lee et al. 2006) and *Cryptococcus neoformans* (Casadevall et al. 2003; Steenbergen and Casadevall 2003).

In addition to virulence factors, the *A. fumigatus* genome encodes 20 allergens (Table S13 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec) and 25 proteins displaying significant sequence similarity to known fungal allergens (Table S14 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec), some of which appear to contribute to its pathogenicity (Bowyer and Denning 2007). For example, *A. fumigatus* Asp f6 (AFUA_1G14550), also known as Mn^{2+}-dependent superoxide dismutase (MnSOD), is specifically recognized by IgE from patients with allergic bronchopulmonary aspergillosis (ABPA) and is differentially expressed during germination (Schwienbacher et al. 2005). The broad distribution of allergens among fungal taxa (Text S1) suggests that *A. fumigatus* possesses the same allergen complement as most other aspergilli and that its effect on hypersensitive individuals can be explained mostly by its ubiquity in the environment.

Our analysis has demonstrated that, similar to known virulence-associated genes, most sexual development genes appear to be under negative (purifying) selection in both sexual and asexual *Aspergillus* species (Text S1 and Table S15 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18404212#supplementary-material-sec). More detailed analysis has revealed four genes in the *N. fischeri* lineage that may be under positive selection. This suggests that a few amino acid changes may enable sexuality in *N. fischeri.* The conservation of sex genes in asexual species is due to a latent sexuality, a recent loss of sexuality, pleiotropy, or parasexual recombination following heterokaryon formation as suggested previously (Dyer and Paoletti 2005; Paoletti et al. 2005).

7.6 Conclusion

Lineage-specific (LS) genes (i.e. genes with limited phylogenetic distribution of orthologs in related species) have been the focal point of many comparative genomic studies, because of the assumption that they may be responsible for phenotypic differences among species and niche adaptation. Our analyses of the genomes of *A. fumigatus* and the two closely related species, *N. fischeri and A. clavatus*, demonstrates that *A. fumigatus* may possess genetic determinants that allow it to establish a successful *in vivo* infection. LS genes that have no orthologs in the other two species comprise 8,5% of the *A. fumigatus* genome and often have accessory functions such as carbohydrate and amino acid metabolism, transport, detoxification, or secondary metabolite biosynthesis. Further analysis showed that these genes have distinct features (e.g. the small gene length and number of introns) and tend to cluster in subtelomeric genomic islands, which may function as "gene dumps/factories". The phylogenies of LS genes, their subtelomeric bias and size differences are consistent with the DDL hypothesis stating that duplication being the primary genetic mechanism responsible for the origin of species-specific genes. The presence of genomic islands indicates that *A. fumigatus* and may possess sophisticated genetic mechanisms that facilitate its adaptation to heterogeneous environments such as soil or a living host.

Chapter 8

Discussion

During my thesis research I came to the conclusion that the genomes of filamentous fungi have some unique evolutionary features not seen in other eukaryotes. Gene clustering, particularly for secondary metabolite synthesis is no doubt the most important of these features.

Together with plants and bacteria, filamentous fungi are an important source of secondary metabolites. While some secondary metabolites are shown to be toxic for human, animal, and plants, many others have been shown to be beneficial for human health.

In my work I encountered three main mechanisms for the creation of a SM gene cluster. First, assembly denovo; this scenario was at the base of the creation of the fumonisin cluster in *Fusarium verticillioides* (Chapter 2). The genes in the pathway were ancestrally scattered and selection acted to re-locate the genes into one locus. The second scenario, which the Ace1 cluster in *Magnaporthe grisea* supports, is the duplication of an ancestral cluster creating two copies, one of which can give rise to a new metabolite (Chapter 4). The third is the horizontal gene transfer scenario. I showed in Chapter 4 that part of the Ace1 cluster in an ancestor of *Magnaporthe grisea* was transferred into the *Aspergillus clavatus* genome.

In addition, I found that the creation of a cluster, in any of the three forms mentioned above, is accompanied with: sequence divergence at the gene level; duplication; and loss and gain of some genes in the cluster. The clustering of secondary metabolite biosynthesis genes is so frequent that I was able to create a program SMURF (Chapter 3) that detects potential secondary metabolite clusters in all sequenced filamentous fungal genomes.

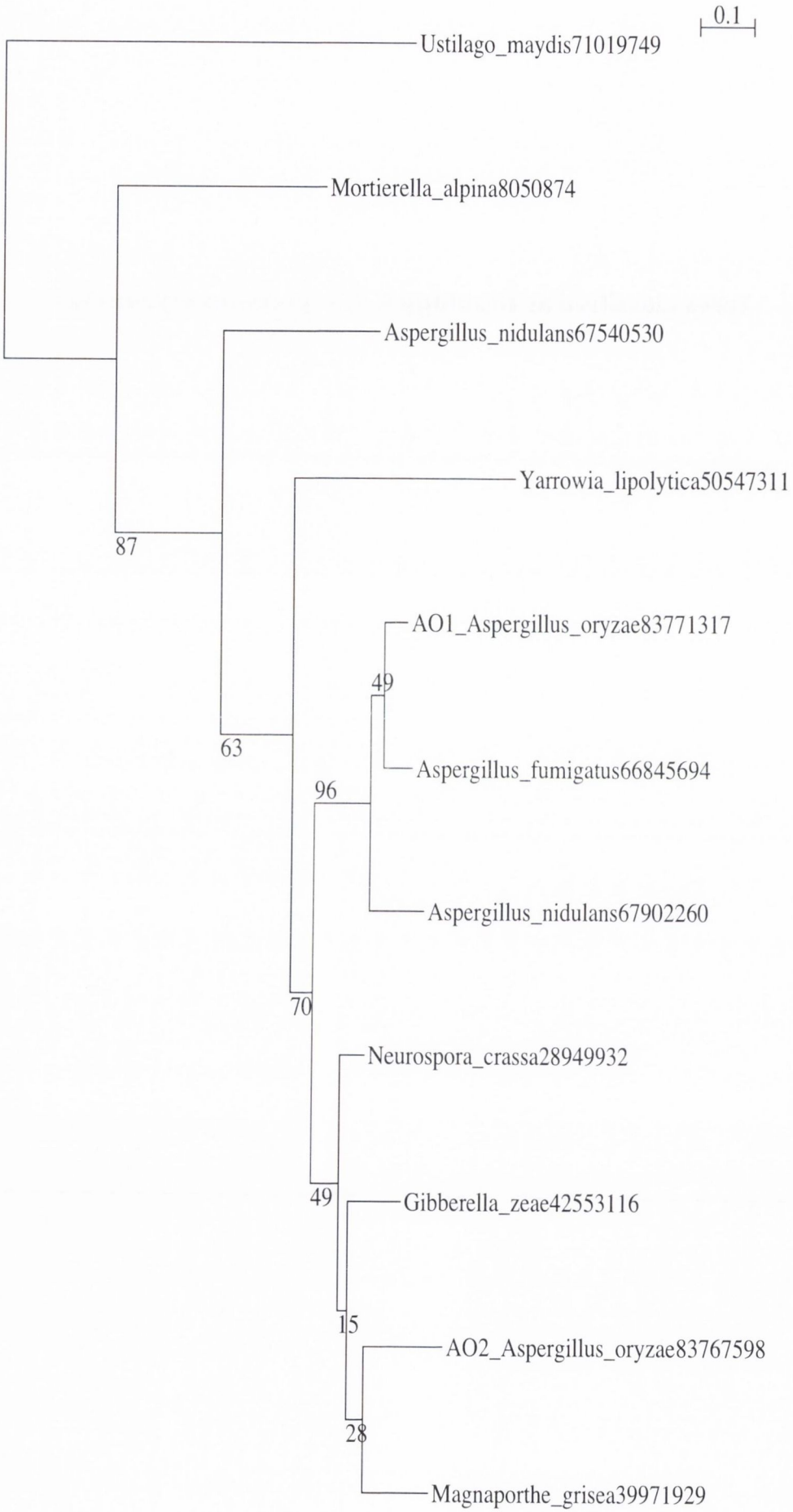
Besides the complex dynamics underlining the gain of secondary metabolite clusters such as duplications and rearrangements. I also observed significant numbers of losses of secondary metabolite clusters. This turnover of secondary metabolite clusters results in important levels of polymorphism even between strains of the same species. For example the strain Af293 from *Aspergillus fumigatus* possesses a unique secondary metabolite cluster, not even found in the CEA10 strain.

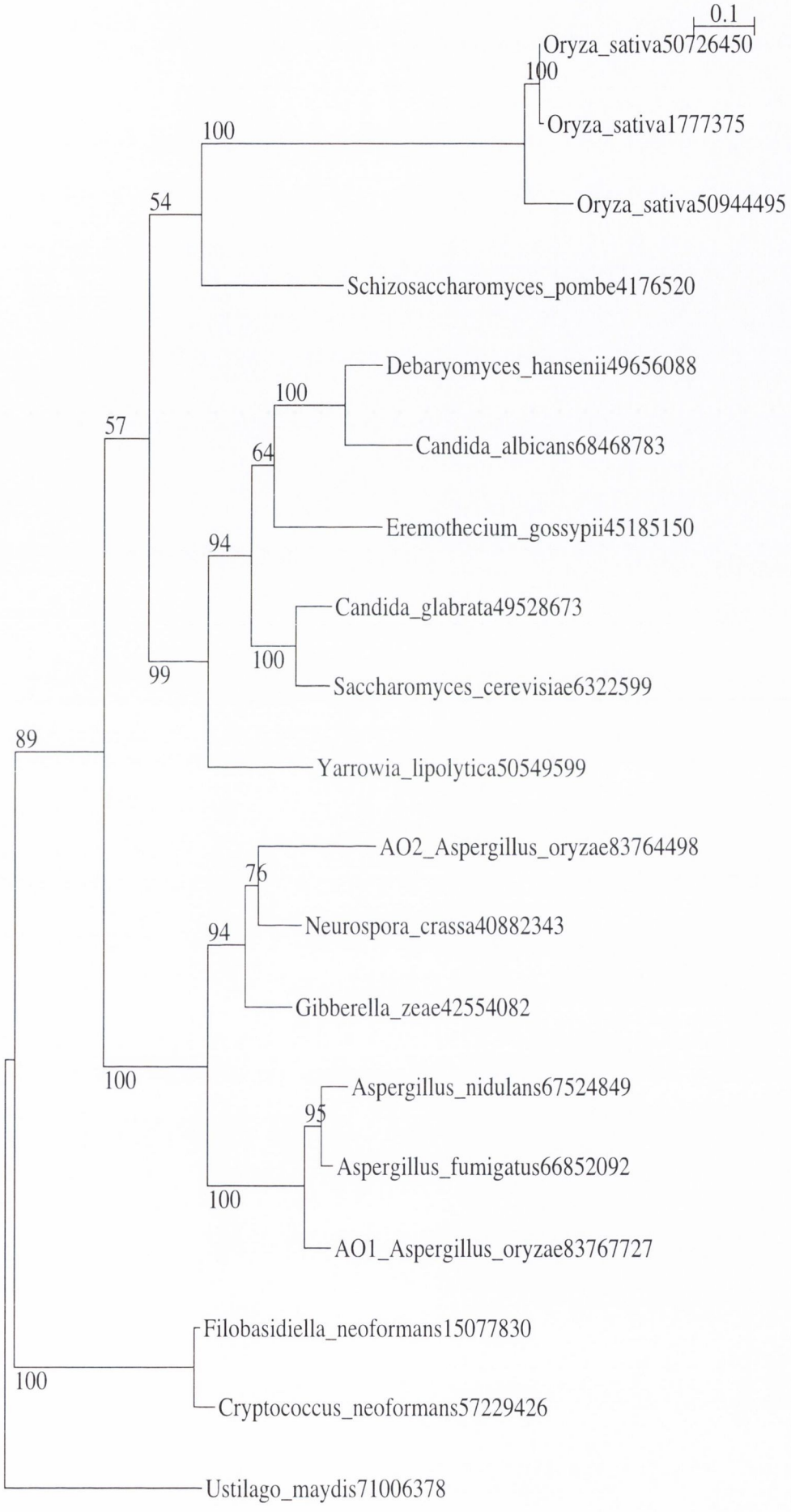
Further to clustering of secondary metabolite genes, another feature that I observed in my thesis research is the high variability of gene content among closely related *Aspergillus* species. A closer analysis of the extra genes in *Aspergillus oryzae* showed that they may be explained in part by frequent horizontal gene transfers from multiple donors, and that there are preferred integration sites of foreign genes in *Aspergillus oryzae*.

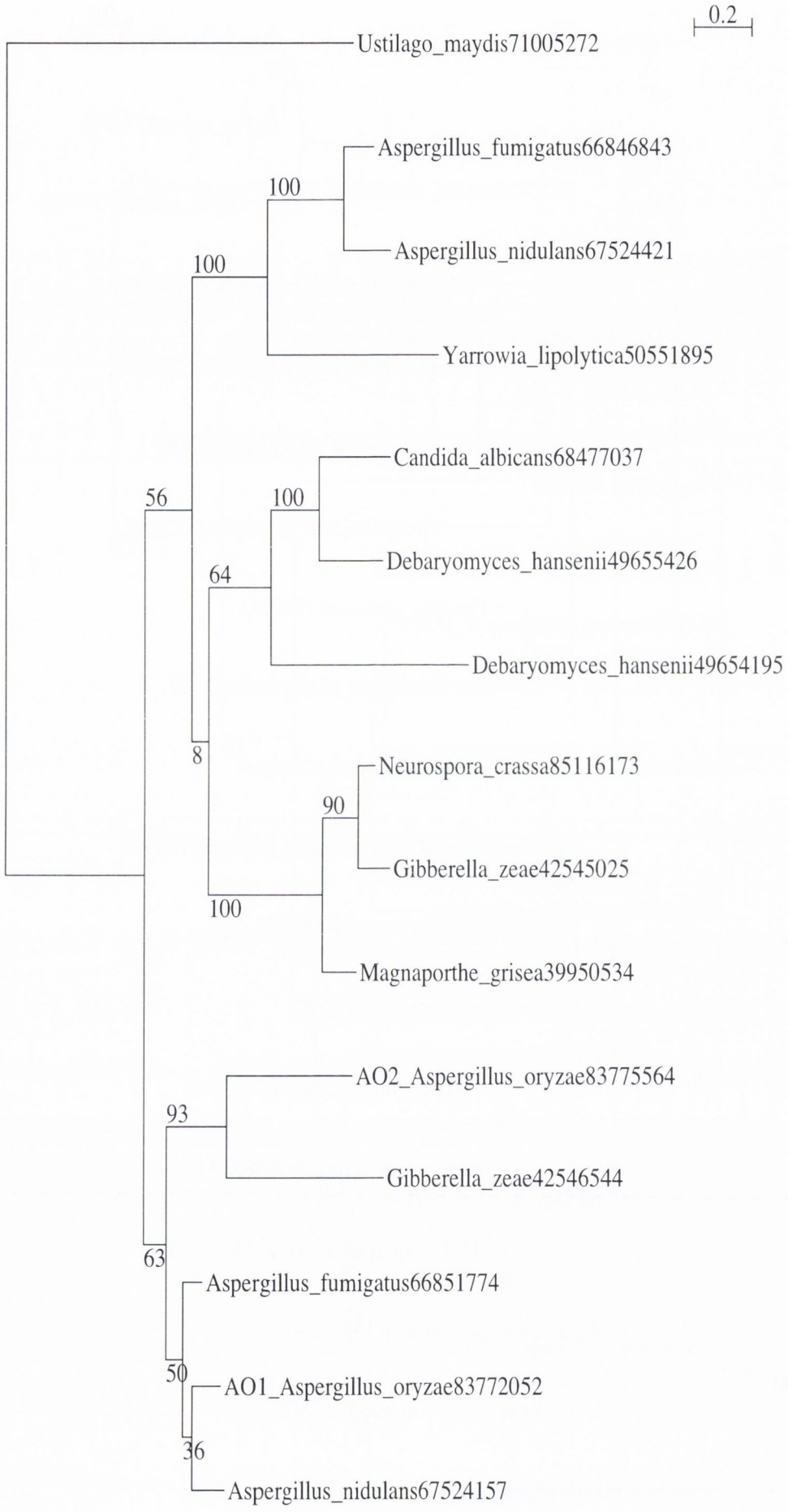
Appendix A

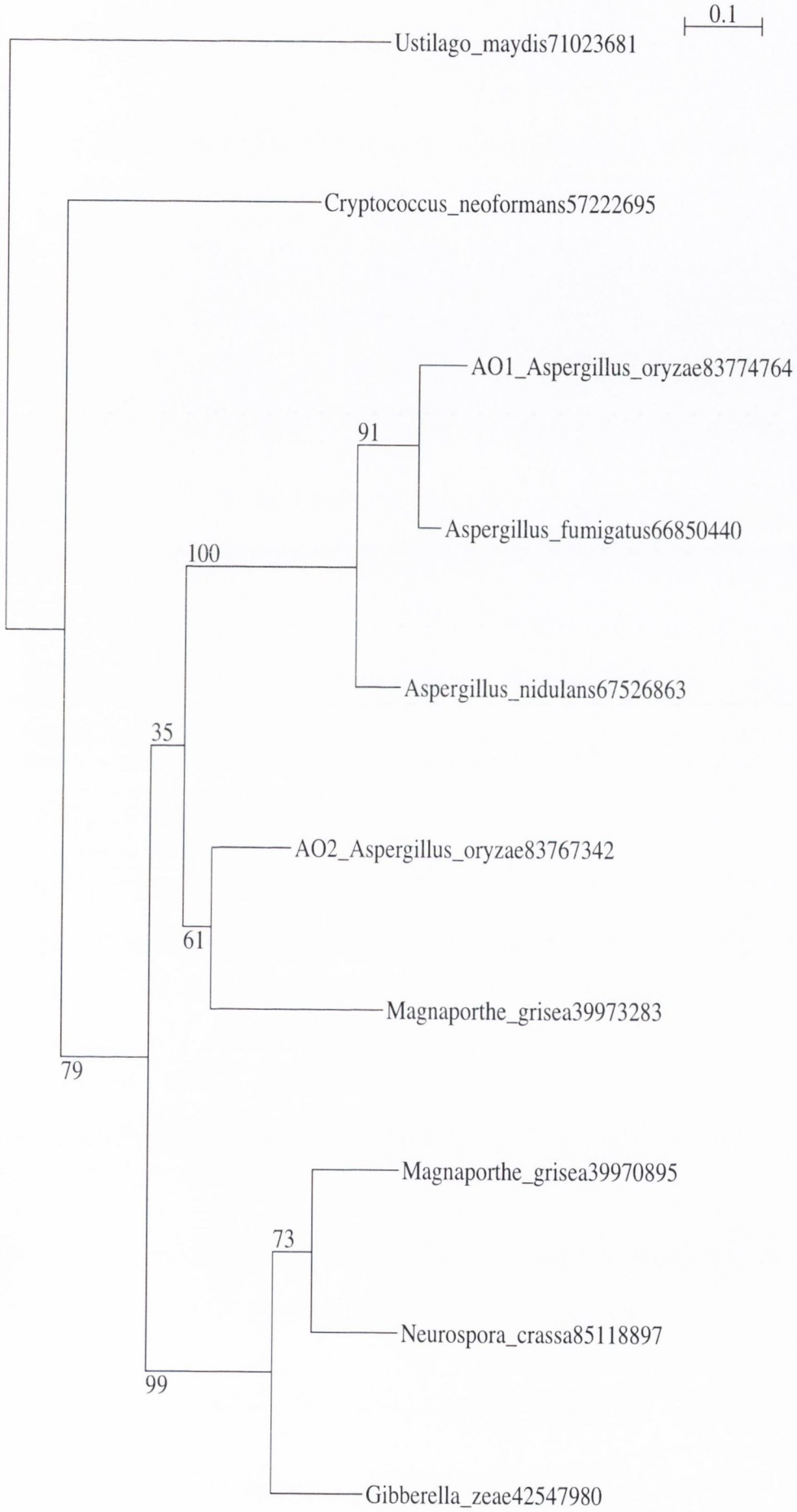
The following section comprises a series of phylogenetic trees relating to Chapter 5 "Elusive Origins of the Extra Genes in *Aspergillus oryzae*".
The trees are classified as Types A, B and C in each *Aspergillus* species. Trees were constructed using PHYML as described in Methods 5.3. In each tree, the sequences identified as *AO1* and *AO2* (for duplications *A. oryzae*), *AN1* and *AN2* (for duplications in *A. nidulans*), or *AFU1* and *AFU2* (for duplications in *A. fumigatus*) are labeled. NCBI identifier (GI) numbers for each sequence are shown.

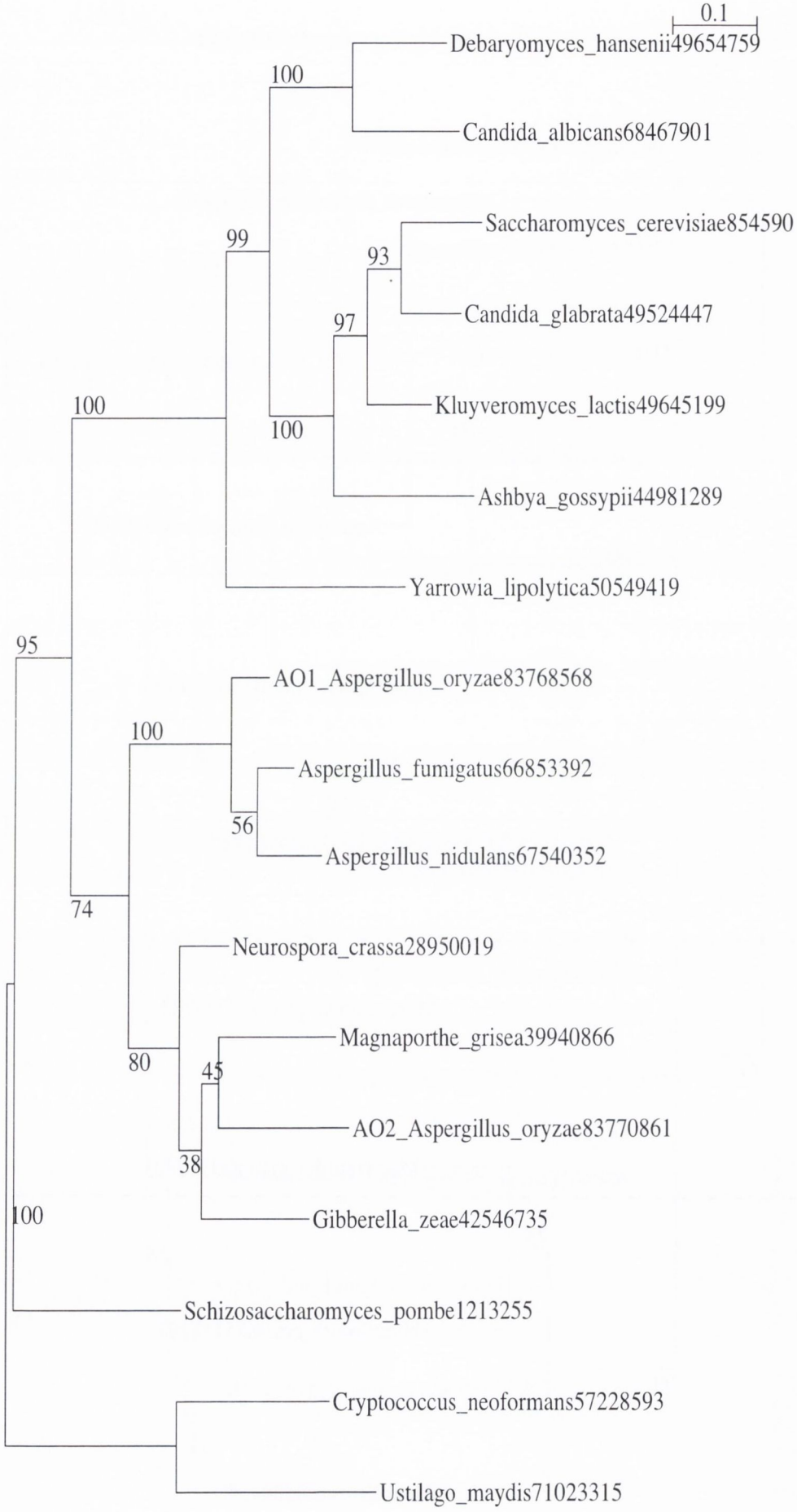
Trees classified as Topology A in *Aspergillus oryzae* (12).

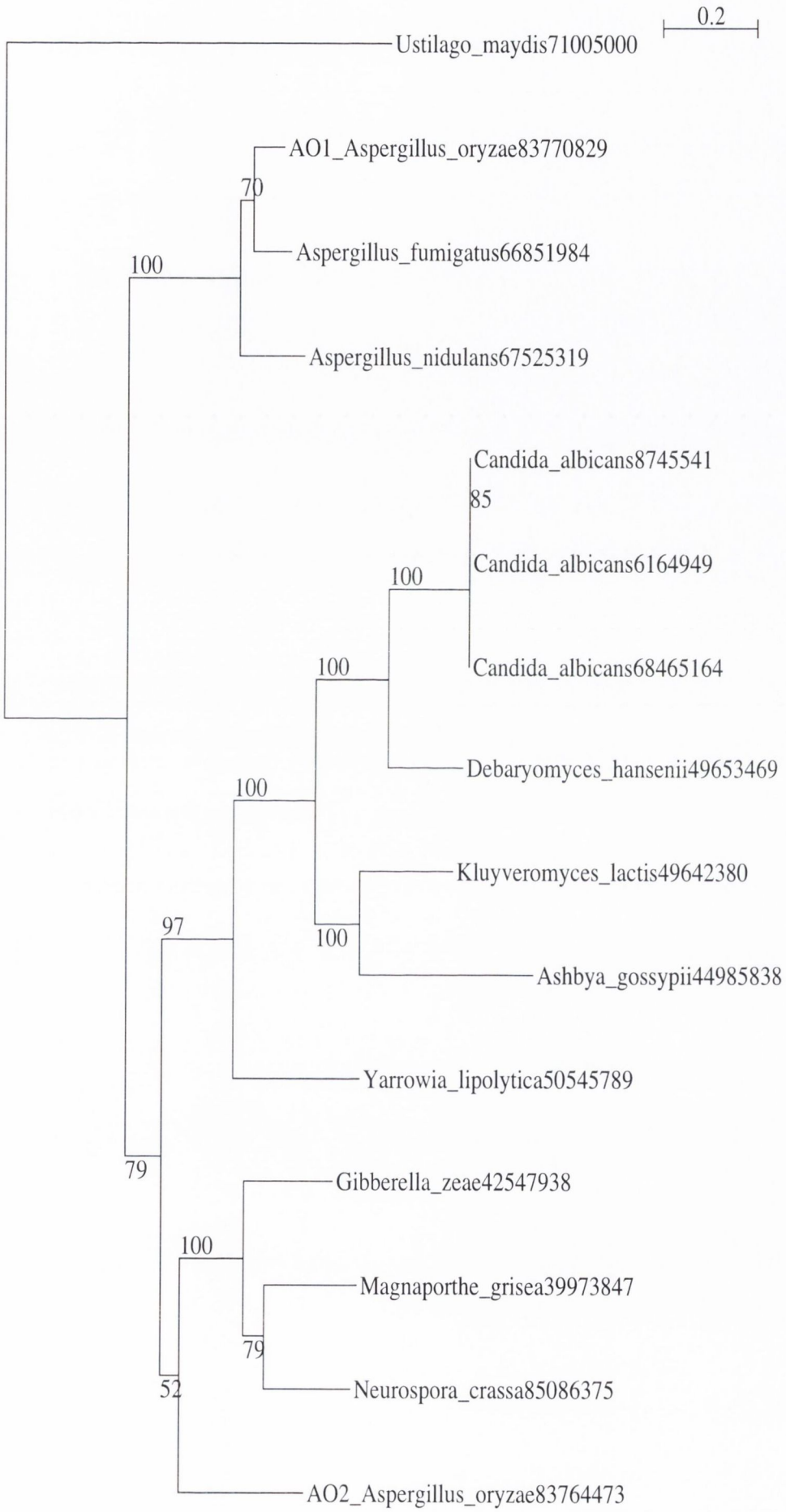

0.1
Ustilago_maydis71019749
Mortierella_alpina8050874
Aspergillus_nidulans67540530
Yarrowia_lipolytica50547311
87
AO1_Aspergillus_oryzae83771317
49
63
Aspergillus_fumigatus66845694
96
Aspergillus_nidulans67902260
70
Neurospora_crassa28949932
49
Gibberella_zeae42553116
15
AO2_Aspergillus_oryzae83767598
28
Magnaporthe_grisea39971929

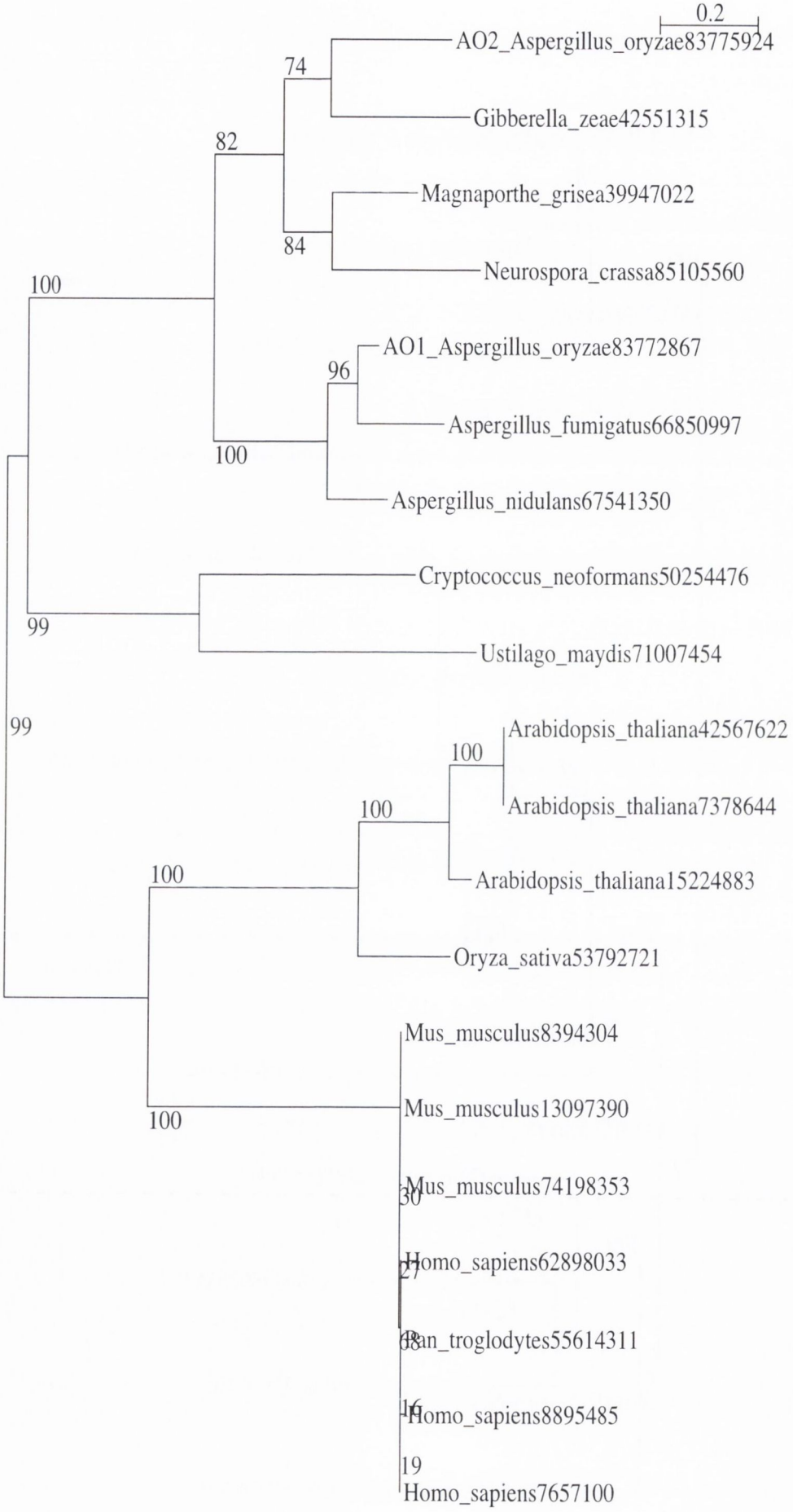

0.1
Oryza_sativa50726450
100
Oryza_sativa1777375
100
Oryza_sativa50944495
54
Schizosaccharomyces_pombe4176520
Debaryomyces_hansenii49656088
100
57
Candida_albicans68468783
64
Eremothecium_gossypii45185150
94
Candida_glabrata49528673
100
99
Saccharomyces_cerevisiae6322599
89
Yarrowia_lipolytica50549599
AO2_Aspergillus_oryzae83764498
76
Neurospora_crassa40882343
94
Gibberella_zeae42554082
100
Aspergillus_nidulans67524849
95
Aspergillus_fumigatus66852092
100
AO1_Aspergillus_oryzae83767727
Filobasidiella_neoformans15077830
100
Cryptococcus_neoformans57229426
Ustilago_maydis71006378

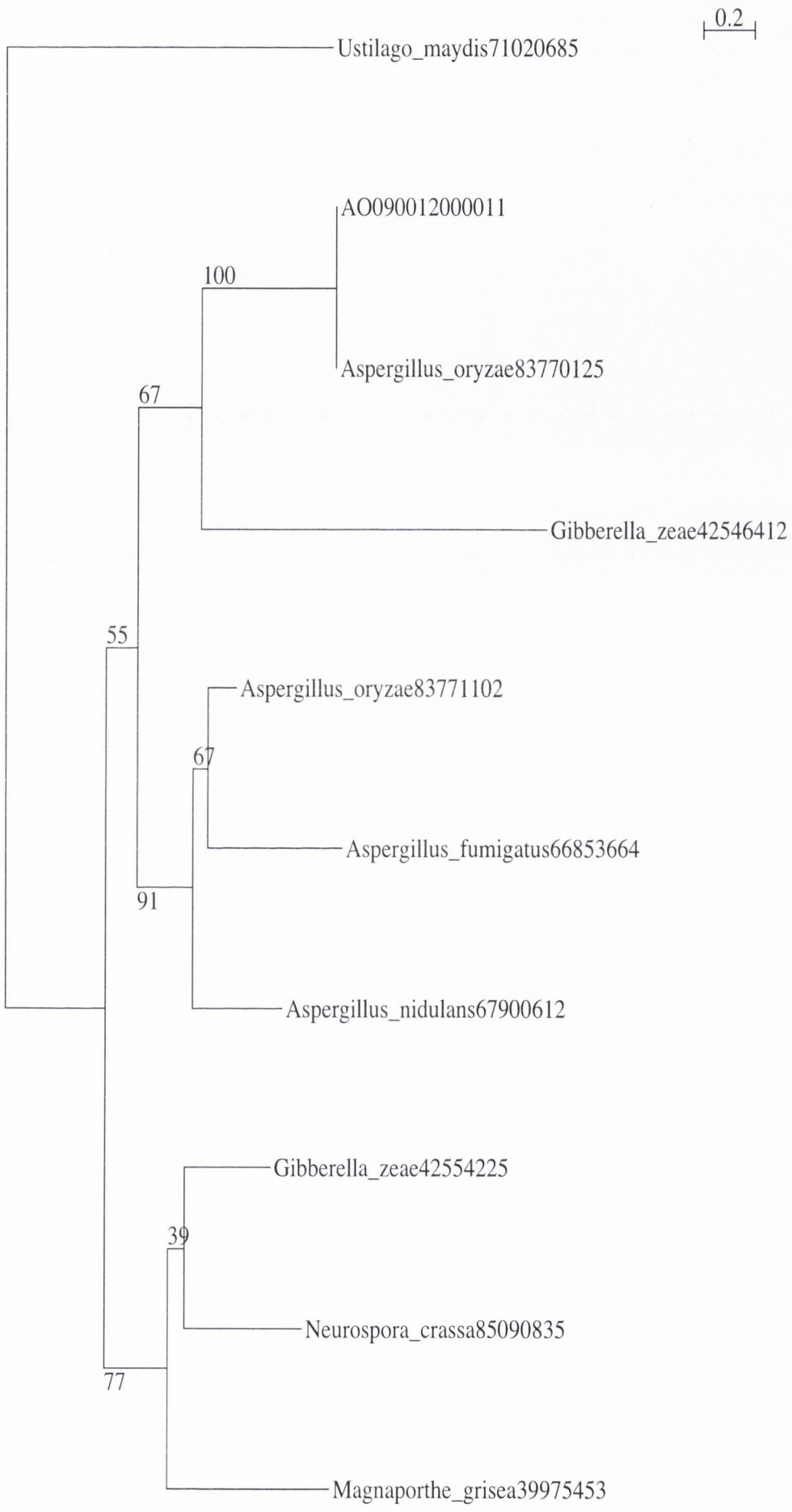

0.2
Ustilago_maydis71005272
Aspergillus_fumigatus66846843
100
Aspergillus_nidulans67524421
100
Yarrowia_lipolytica50551895
Candida_albicans68477037
56
100
Debaryomyces_hansenii49655426
64
Debaryomyces_hansenii49654195
8
Neurospora_crassa85116173
90
Gibberella_zeae42545025
100
Magnaporthe_grisea39950534
AO2_Aspergillus_oryzae83775564
93
Gibberella_zeae42546544
63
Aspergillus_fumigatus66851774
50
AO1_Aspergillus_oryzae83772052
36
Aspergillus_nidulans67524157

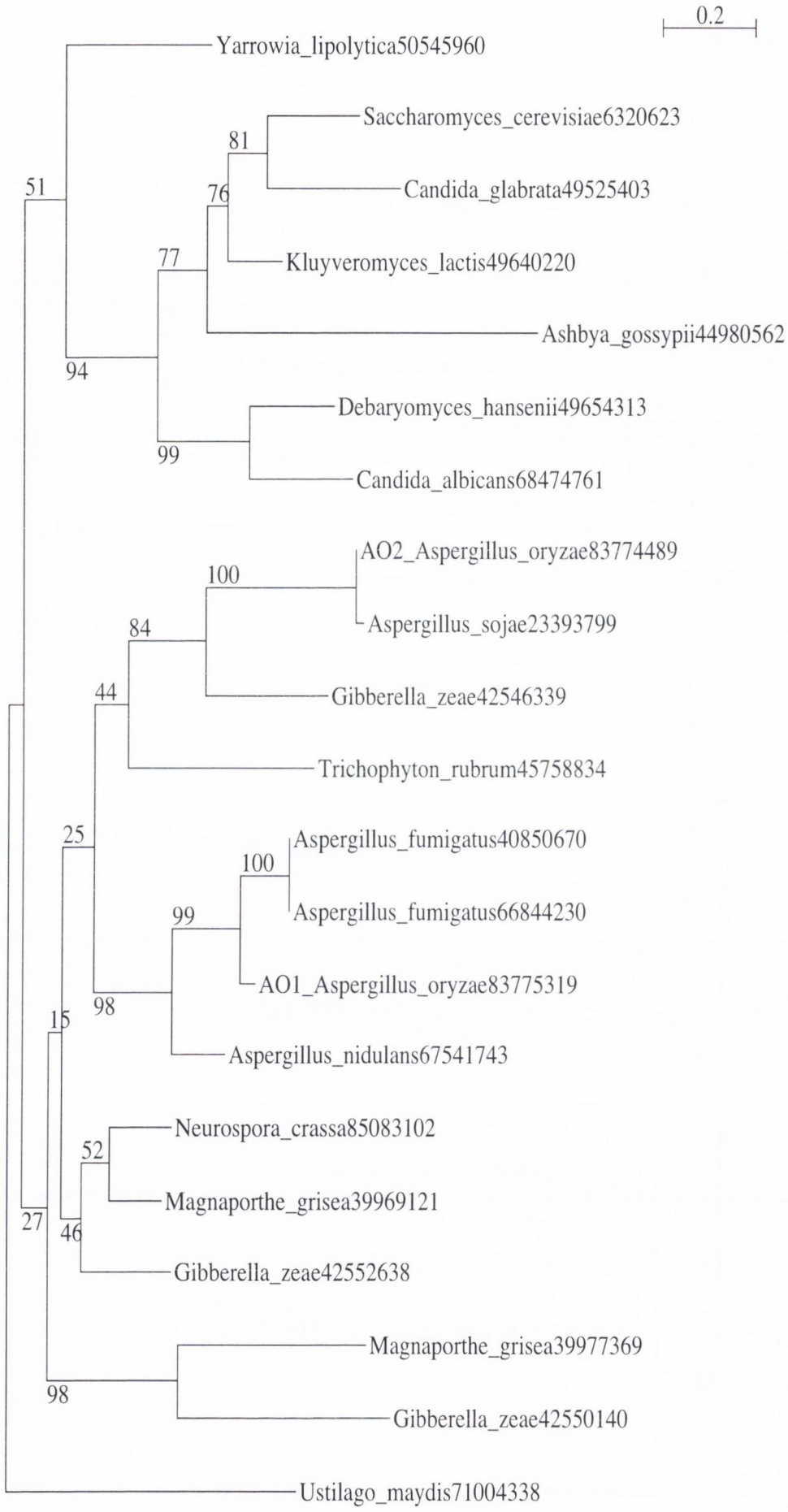

0.1
Ustilago_maydis71023681
Cryptococcus_neoformans57222695
AO1_Aspergillus_oryzae83774764
91
Aspergillus_fumigatus66850440
100
Aspergillus_nidulans67526863
35
AO2_Aspergillus_oryzae83767342
61
Magnaporthe_grisea39973283
79
Magnaporthe_grisea39970895
73
Neurospora_crassa85118897
99
Gibberella_zeae42547980

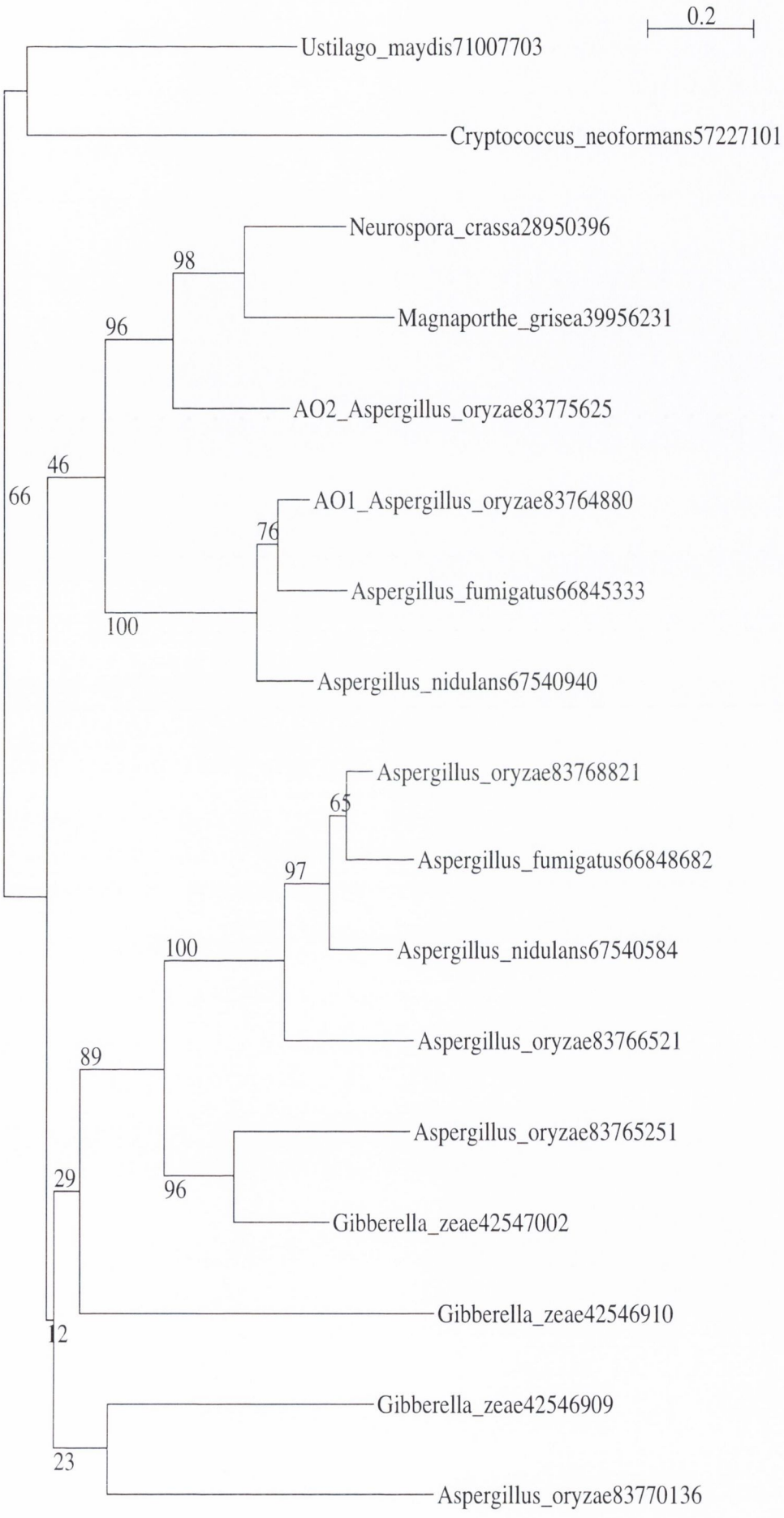

0.1
Debaryomyces_hansenii49654759
Candida_albicans68467901
Saccharomyces_cerevisiae854590
Candida_glabrata49524447
Kluyveromyces_lactis49645199
Ashbya_gossypii44981289
Yarrowia_lipolytica50549419
AO1_Aspergillus_oryzae83768568
Aspergillus_fumigatus66853392
Aspergillus_nidulans67540352
Neurospora_crassa28950019
Magnaporthe_grisea39940866
AO2_Aspergillus_oryzae83770861
Gibberella_zeae42546735
Schizosaccharomyces_pombe1213255
Cryptococcus_neoformans57228593
Ustilago_maydis71023315
100
99
93
97
100
100
95
100
56
74
80
45
38
100

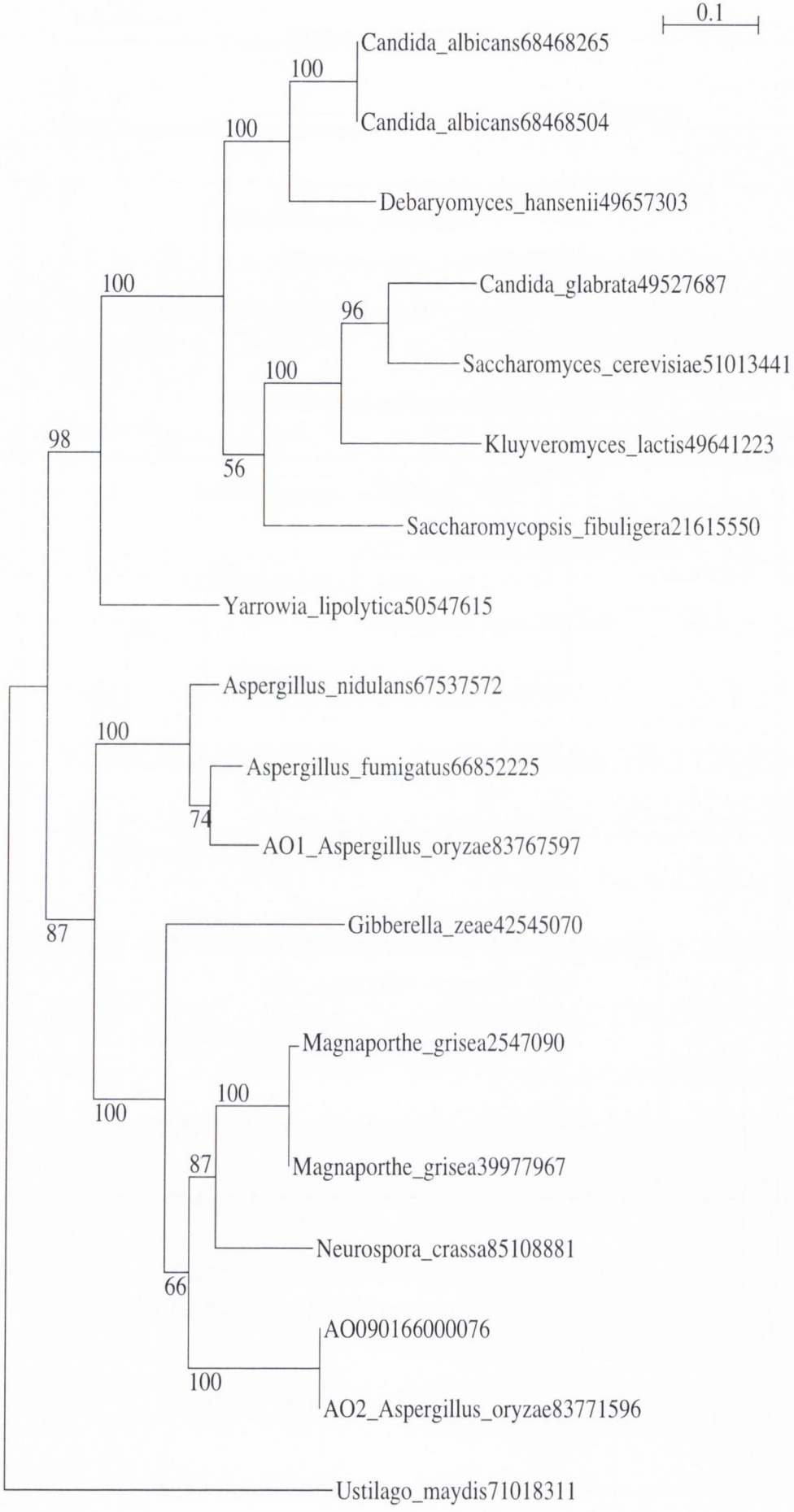

0.2
Ustilago_maydis71005000
AO1_Aspergillus_oryzae83770829
70
Aspergillus_fumigatus66851984
100
Aspergillus_nidulans67525319
Candida_albicans8745541
85
Candida_albicans6164949
100
Candida_albicans68465164
100
Debaryomyces_hansenii49653469
100
Kluyveromyces_lactis49642380
97
100
Ashbya_gossypii44985838
Yarrowia_lipolytica50545789
79
Gibberella_zeae42547938
100
Magnaporthe_grisea39973847
79
Neurospora_crassa85086375
52
AO2_Aspergillus_oryzae83764473

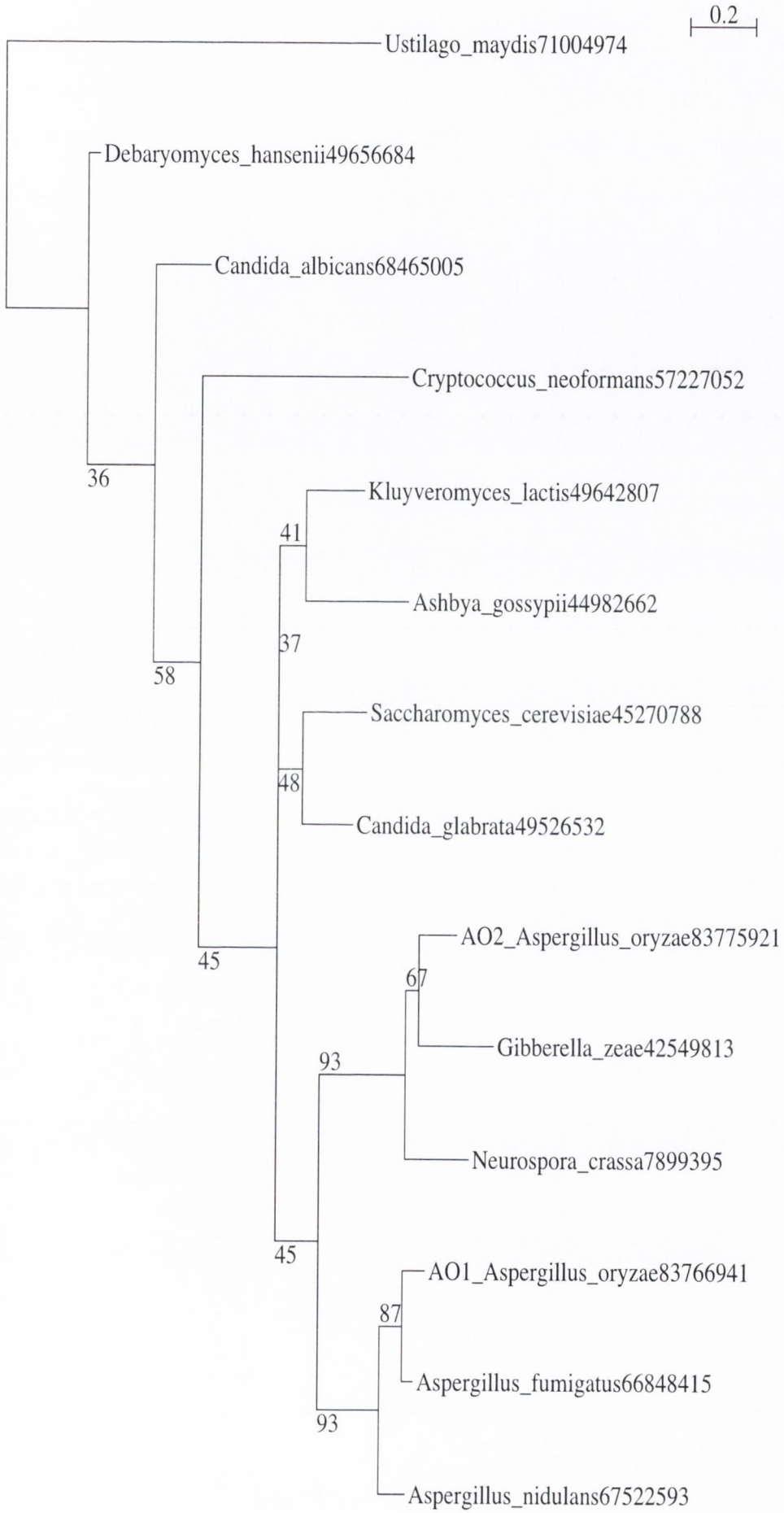

0.2
AO2_Aspergillus_oryzae83775924
Gibberella_zeae42551315
Magnaporthe_grisea39947022
Neurospora_crassa85105560
AO1_Aspergillus_oryzae83772867
Aspergillus_fumigatus66850997
Aspergillus_nidulans67541350
Cryptococcus_neoformans50254476
Ustilago_maydis71007454
Arabidopsis_thaliana42567622
Arabidopsis_thaliana7378644
Arabidopsis_thaliana15224883
Oryza_sativa53792721
Mus_musculus8394304
Mus_musculus13097390
Mus_musculus74198353
Homo_sapiens62898033
Pan_troglodytes55614311
Homo_sapiens8895485
Homo_sapiens7657100
74
82
84
100
96
100
100
99
99
100
100
100
100
30
27
16
19


0.2
Ustilago_maydis71020685
AO090012000011
100
Aspergillus_oryzae83770125
67
Gibberella_zeae42546412
55
Aspergillus_oryzae83771102
67
Aspergillus_fumigatus66853664
91
Aspergillus_nidulans67900612
Gibberella_zeae42554225
39
Neurospora_crassa85090835
77
Magnaporthe_grisea39975453

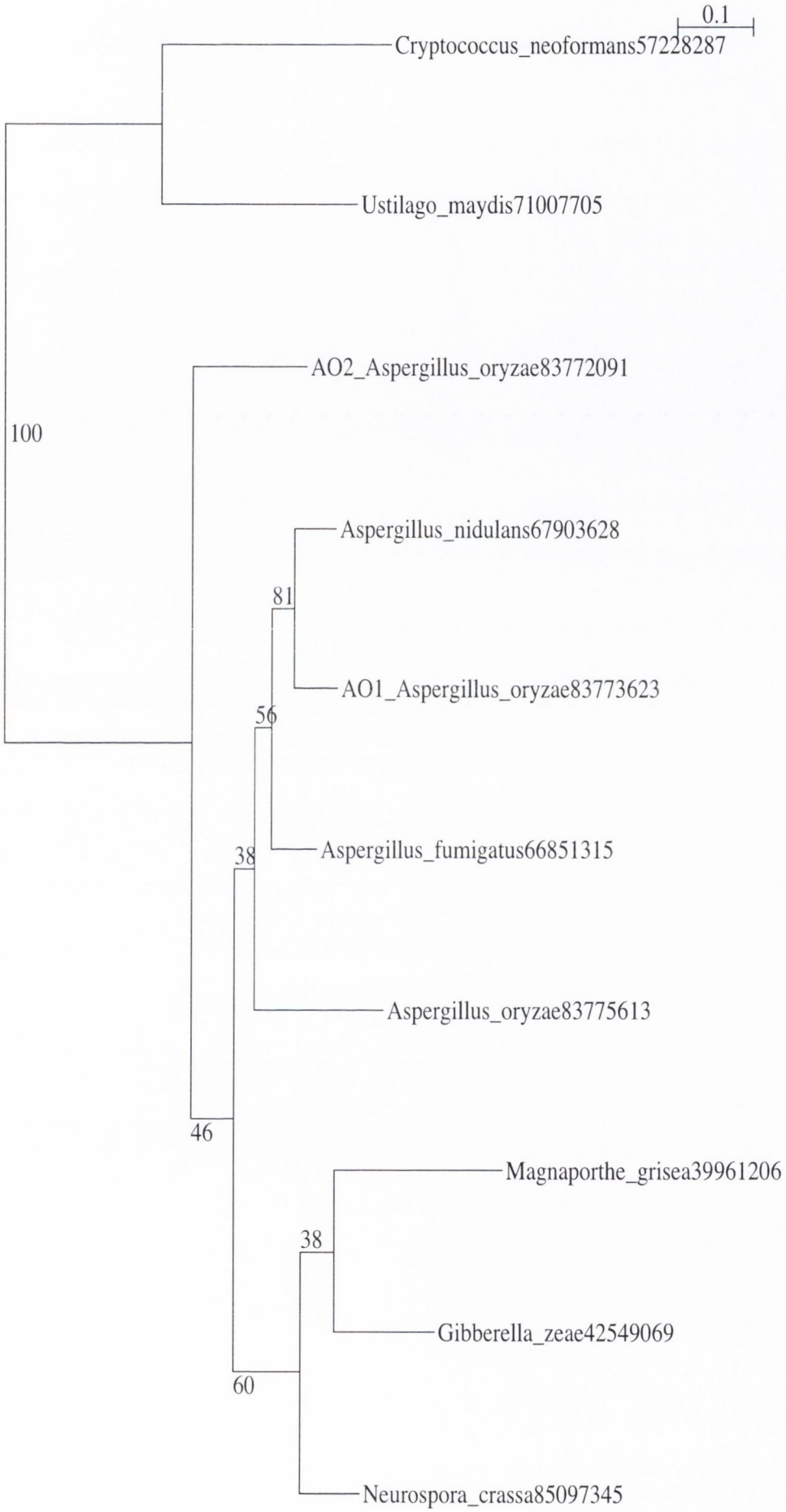

0.2
Yarrowia_lipolytica50545960
Saccharomyces_cerevisiae6320623
Candida_glabrata49525403
Kluyveromyces_lactis49640220
Ashbya_gossypii44980562
Debaryomyces_hansenii49654313
Candida_albicans68474761
AO2_Aspergillus_oryzae83774489
Aspergillus_sojae23393799
Gibberella_zeae42546339
Trichophyton_rubrum45758834
Aspergillus_fumigatus40850670
Aspergillus_fumigatus66844230
AO1_Aspergillus_oryzae83775319
Aspergillus_nidulans67541743
Neurospora_crassa85083102
Magnaporthe_grisea39969121
Gibberella_zeae42552638
Magnaporthe_grisea39977369
Gibberella_zeae42550140
Ustilago_maydis71004338
81
51
76
77
94
99
100
84
44
25
100
99
98
15
52
27
46
98

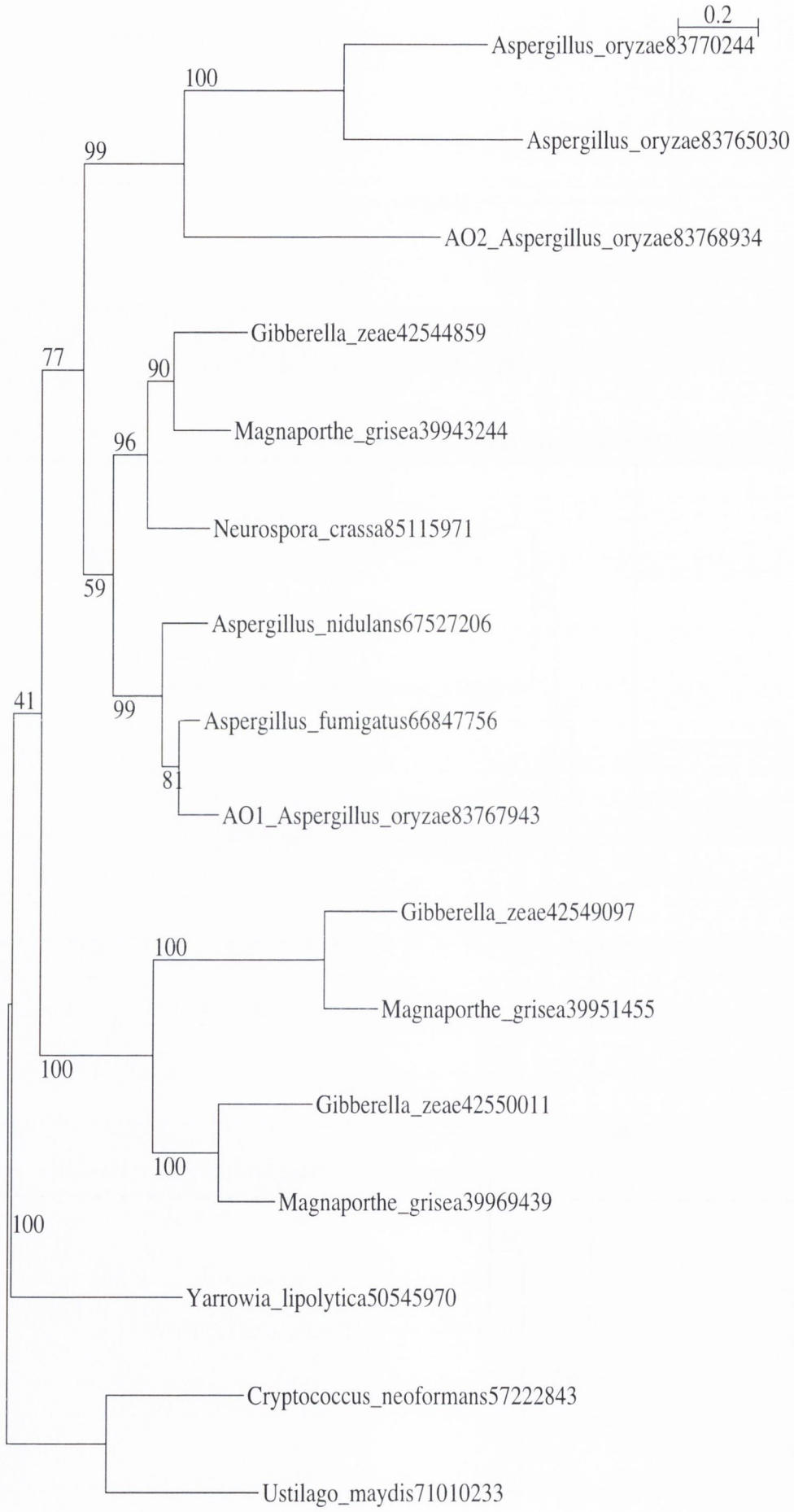

0.2
Ustilago_maydis71007703
Cryptococcus_neoformans57227101
Neurospora_crassa28950396
98
Magnaporthe_grisea39956231
96
AO2_Aspergillus_oryzae83775625
46
66
AO1_Aspergillus_oryzae83764880
76
Aspergillus_fumigatus66845333
100
Aspergillus_nidulans67540940
Aspergillus_oryzae83768821
65
Aspergillus_fumigatus66848682
97
Aspergillus_nidulans67540584
100
Aspergillus_oryzae83766521
89
Aspergillus_oryzae83765251
29
96
Gibberella_zeae42547002
Gibberella_zeae42546910
12
Gibberella_zeae42546909
23
Aspergillus_oryzae83770136

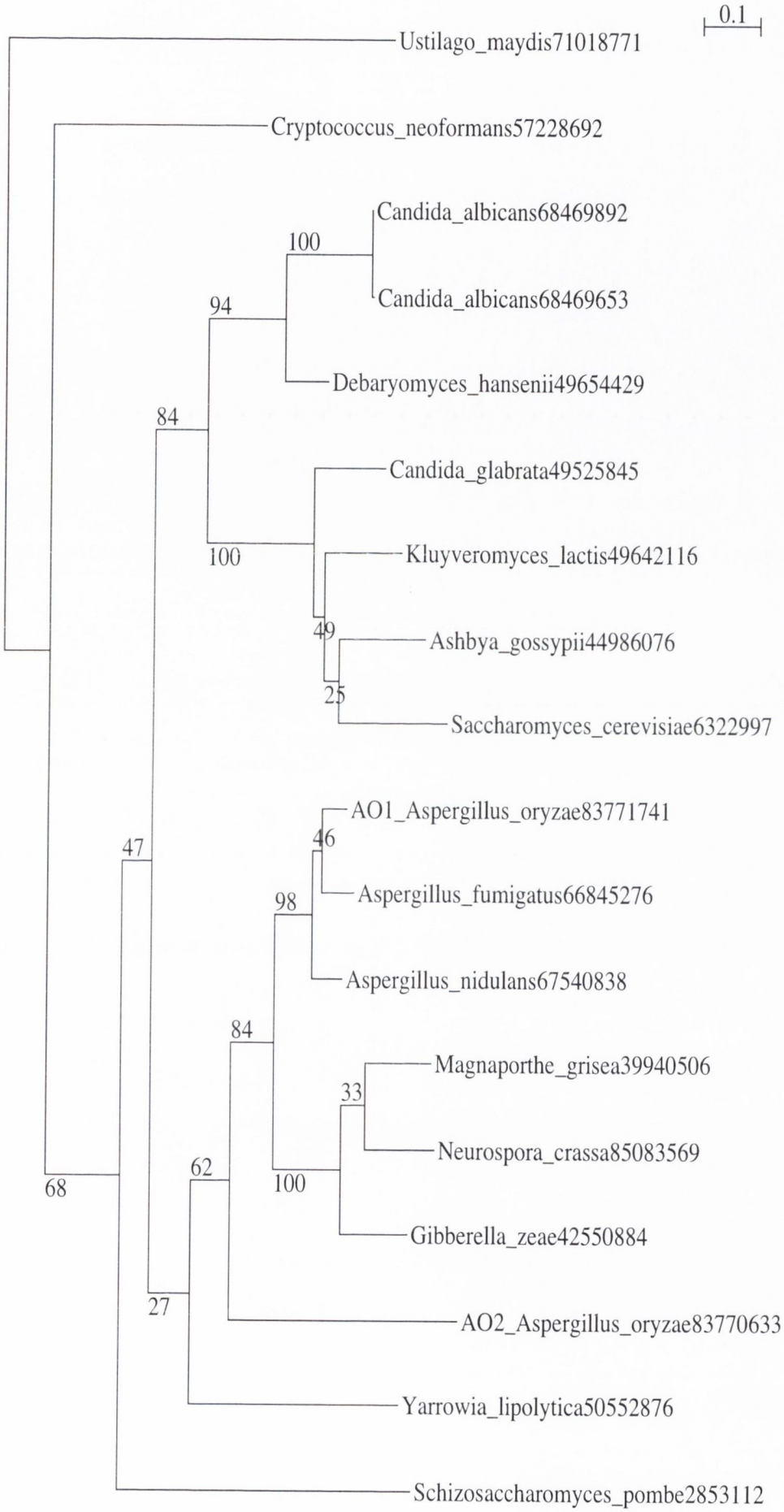

0.1
Candida_albicans68468265
Candida_albicans68468504
Debaryomyces_hansenii49657303
Candida_glabrata49527687
Saccharomyces_cerevisiae51013441
Kluyveromyces_lactis49641223
Saccharomycopsis_fibuligera21615550
Yarrowia_lipolytica50547615
Aspergillus_nidulans67537572
Aspergillus_fumigatus66852225
AO1_Aspergillus_oryzae83767597
Gibberella_zeae42545070
Magnaporthe_grisea2547090
Magnaporthe_grisea39977967
Neurospora_crassa85108881
AO090166000076
AO2_Aspergillus_oryzae83771596
Ustilago_maydis71018311
100
100
100
96
100
98
56
100
74
87
100
100
87
66
100

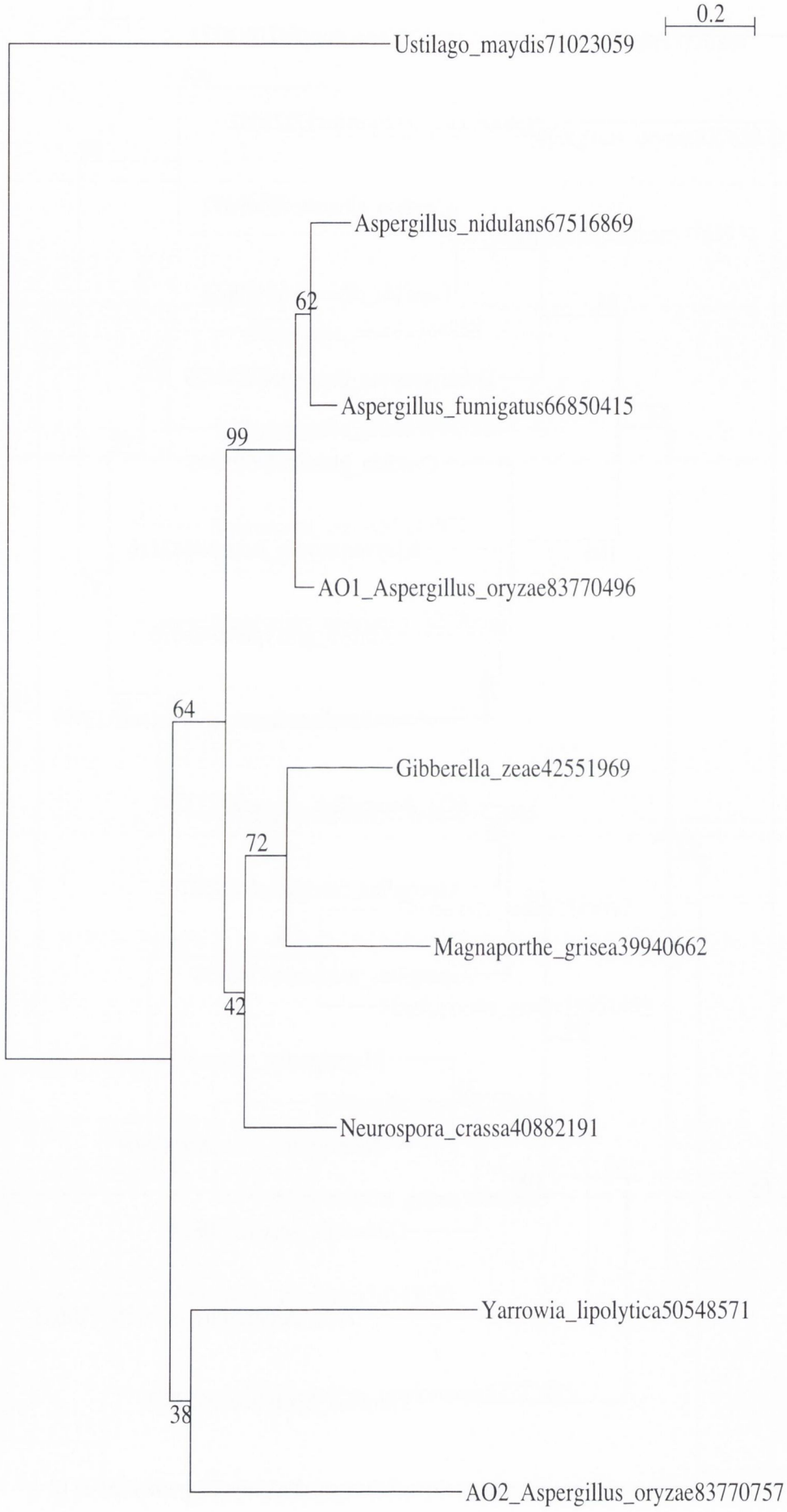

0.2
Ustilago_maydis71004974
Debaryomyces_hansenii49656684
Candida_albicans68465005
Cryptococcus_neoformans57227052
36
Kluyveromyces_lactis49642807
41
Ashbya_gossypii44982662
37
58
Saccharomyces_cerevisiae45270788
48
Candida_glabrata49526532
AO2_Aspergillus_oryzae83775921
45
67
Gibberella_zeae42549813
93
Neurospora_crassa7899395
45
AO1_Aspergillus_oryzae83766941
87
Aspergillus_fumigatus66848415
93
Aspergillus_nidulans67522593

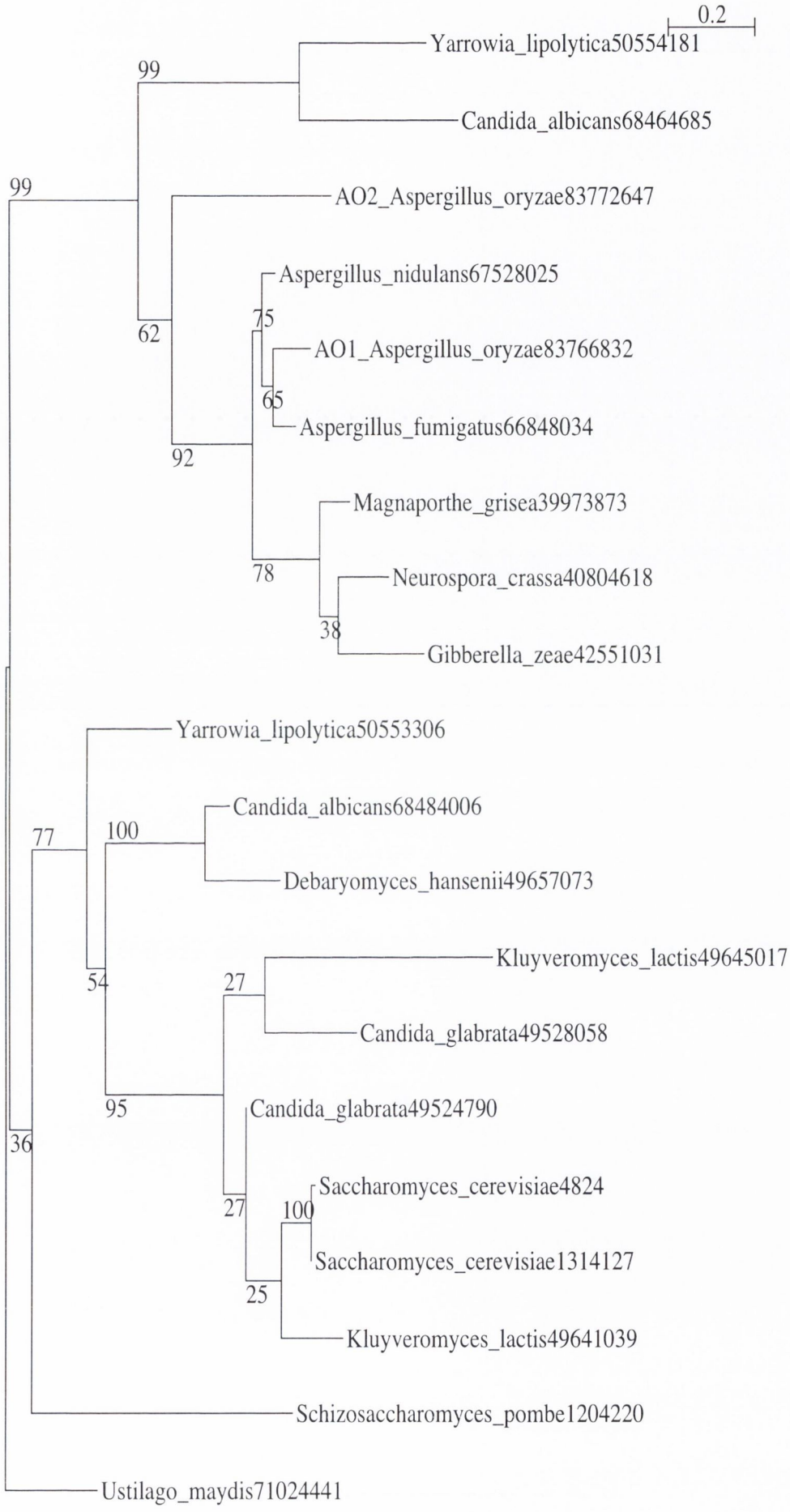
Trees classified as Topology B in *Aspergillus oryzae* (9).

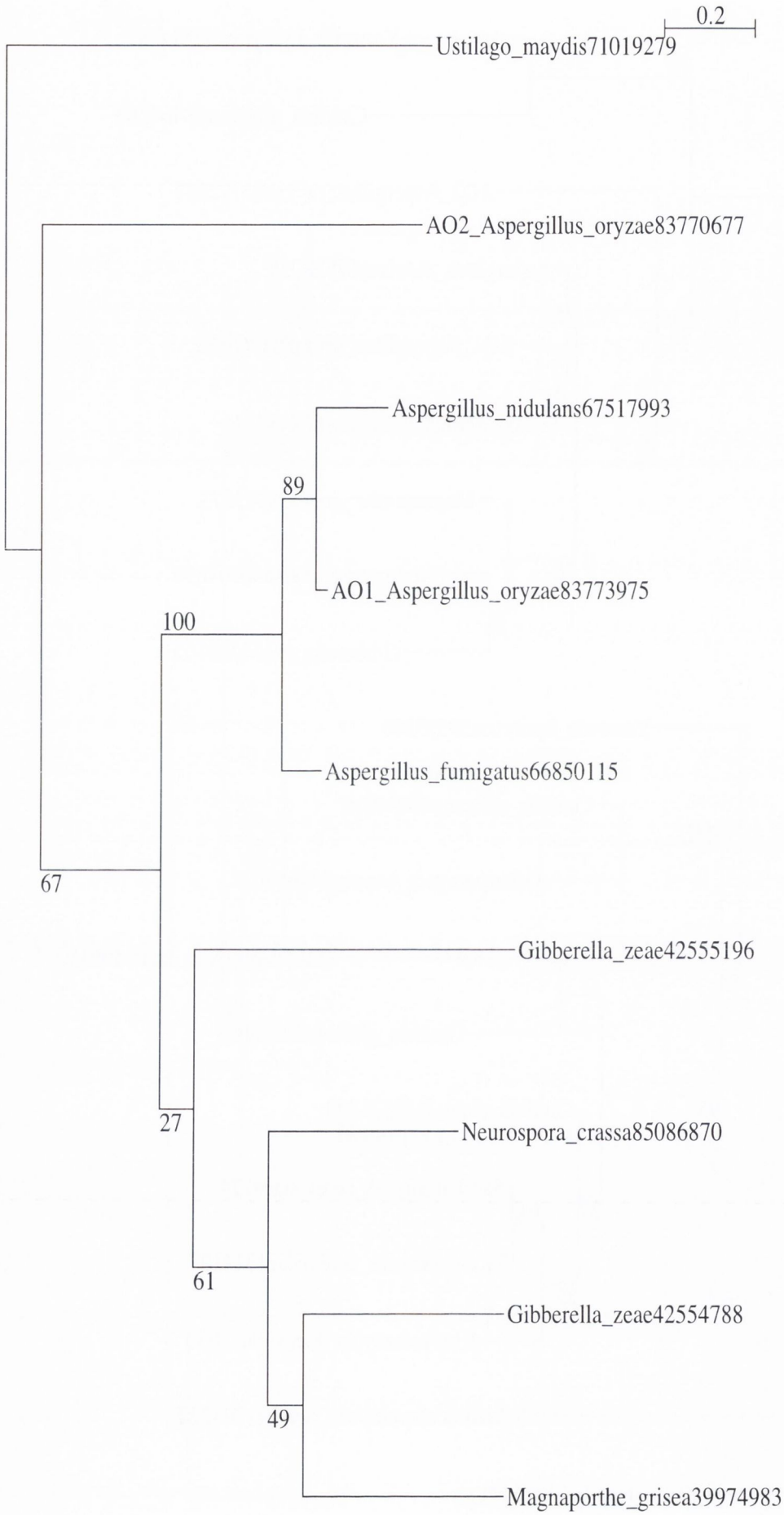

0.1
Cryptococcus_neoformans57228287
Ustilago_maydis71007705
AO2_Aspergillus_oryzae83772091
100
Aspergillus_nidulans67903628
81
AO1_Aspergillus_oryzae83773623
56
Aspergillus_fumigatus66851315
38
Aspergillus_oryzae83775613
46
Magnaporthe_grisea39961206
38
Gibberella_zeae42549069
60
Neurospora_crassa85097345

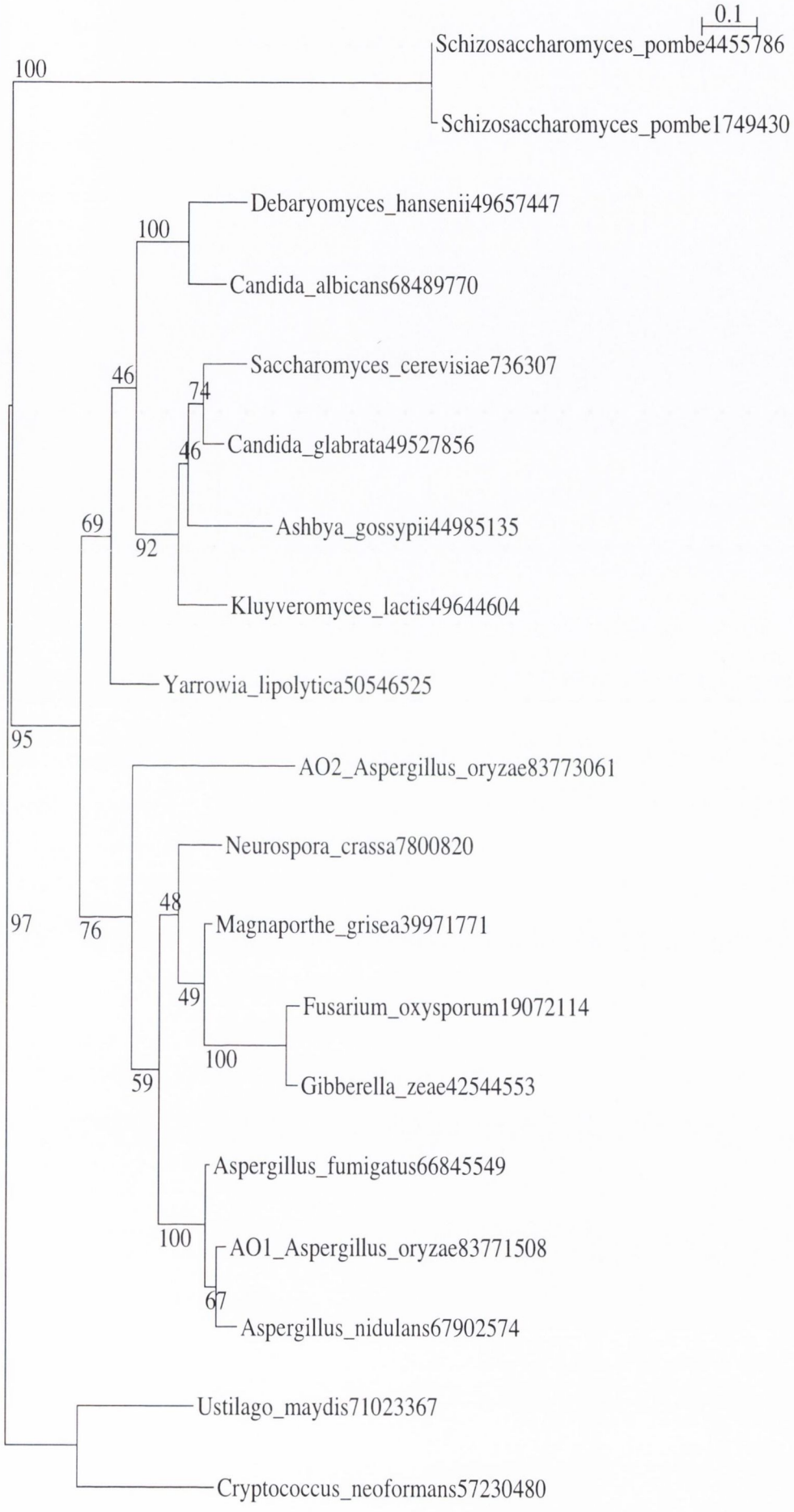

0.2
Aspergillus_oryzae83770244
Aspergillus_oryzae83765030
AO2_Aspergillus_oryzae83768934
Gibberella_zeae42544859
Magnaporthe_grisea39943244
Neurospora_crassa85115971
Aspergillus_nidulans67527206
Aspergillus_fumigatus66847756
AO1_Aspergillus_oryzae83767943
Gibberella_zeae42549097
Magnaporthe_grisea39951455
Gibberella_zeae42550011
Magnaporthe_grisea39969439
Yarrowia_lipolytica50545970
Cryptococcus_neoformans57222843
Ustilago_maydis71010233
100
99
100
77
90
96
59
99
81
41
100
100
100

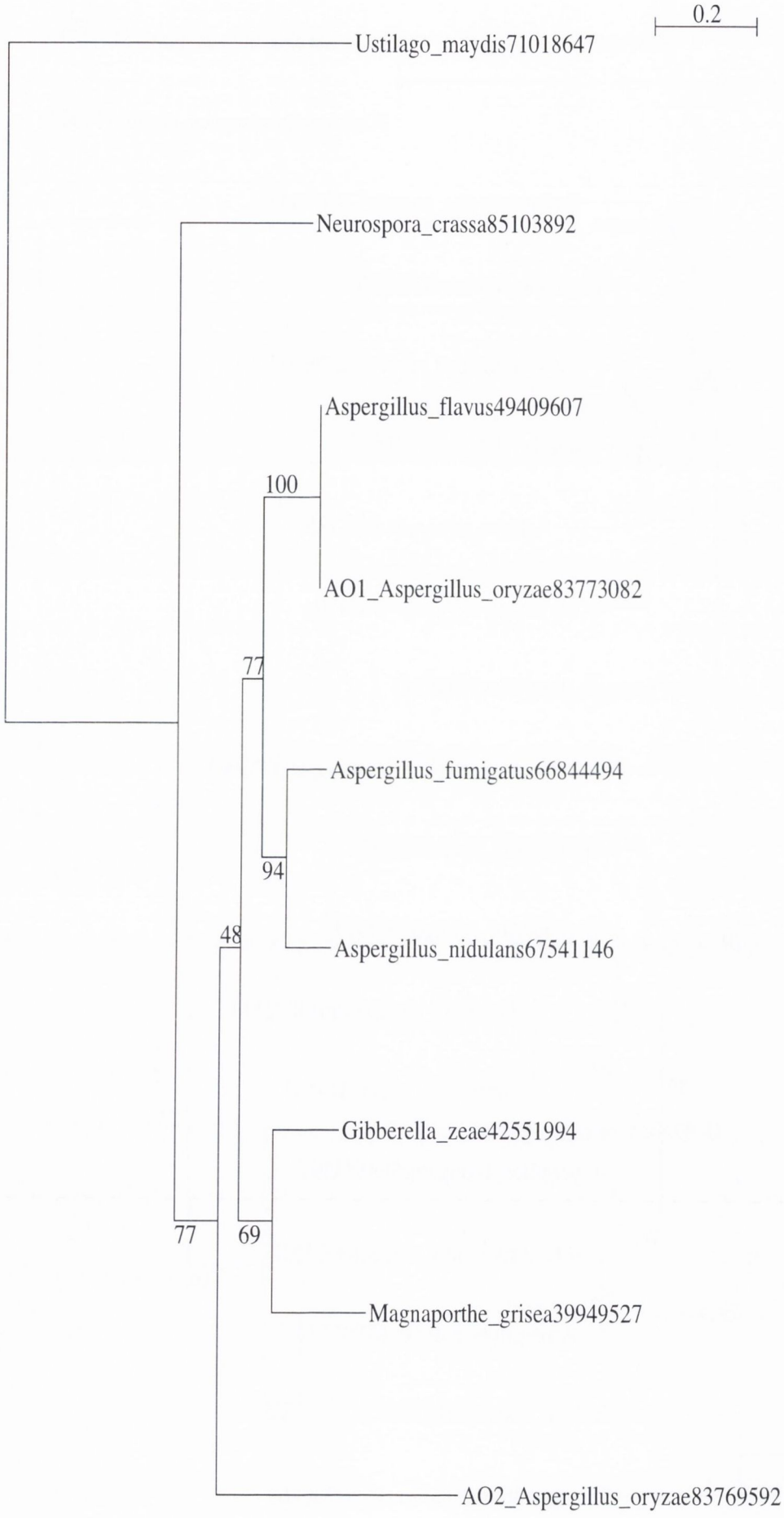

0.1
Ustilago_maydis71018771
Cryptococcus_neoformans57228692
Candida_albicans68469892
Candida_albicans68469653
Debaryomyces_hansenii49654429
Candida_glabrata49525845
Kluyveromyces_lactis49642116
Ashbya_gossypii44986076
Saccharomyces_cerevisiae6322997
AO1_Aspergillus_oryzae83771741
Aspergillus_fumigatus66845276
Aspergillus_nidulans67540838
Magnaporthe_grisea39940506
Neurospora_crassa85083569
Gibberella_zeae42550884
AO2_Aspergillus_oryzae83770633
Yarrowia_lipolytica50552876
Schizosaccharomyces_pombe2853112
100
94
84
100
49
25
47
46
98
84
33
62
100
68
27

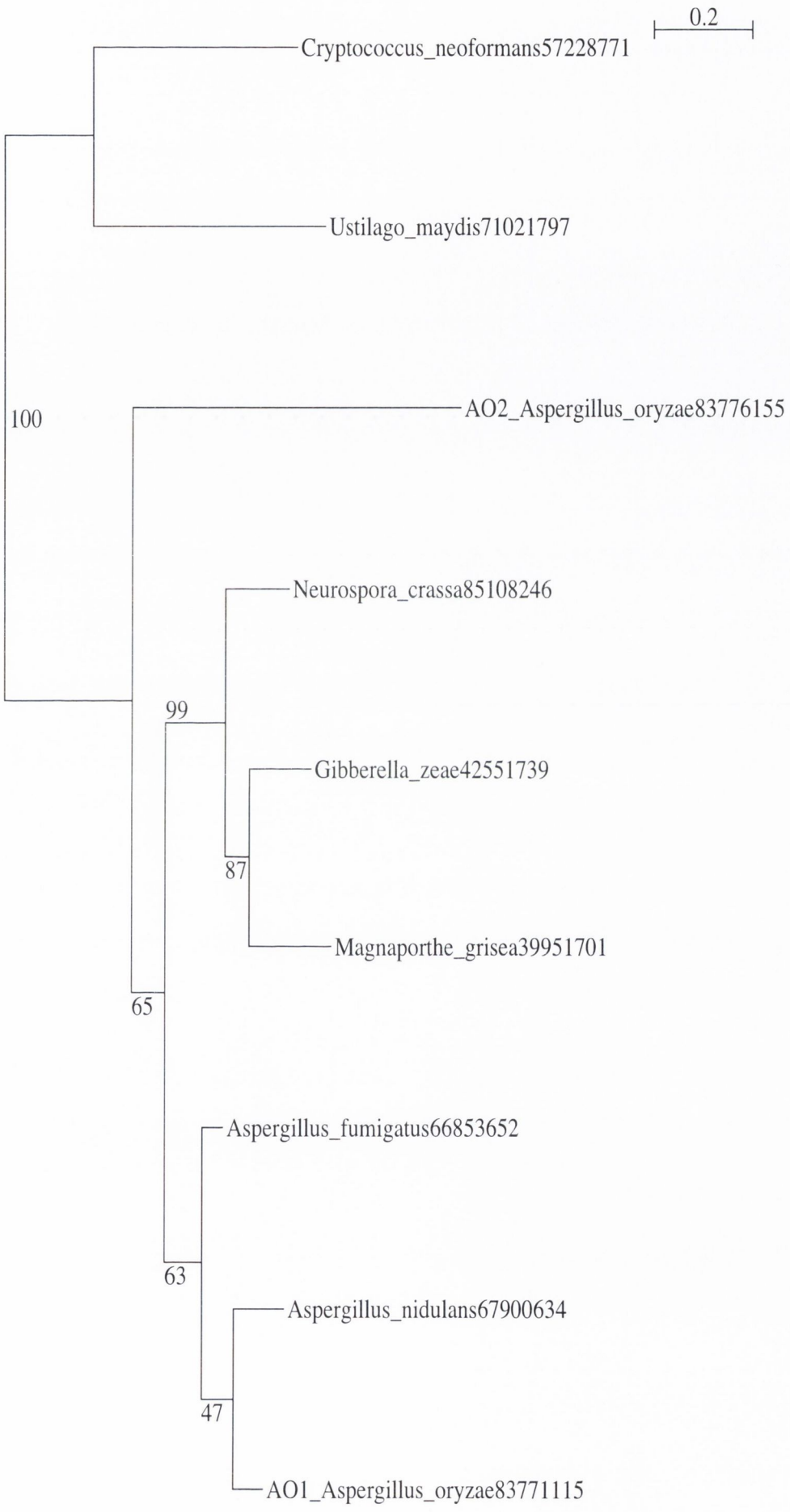

0.2
Ustilago_maydis71023059
Aspergillus_nidulans67516869
62
Aspergillus_fumigatus66850415
99
AO1_Aspergillus_oryzae83770496
64
Gibberella_zeae42551969
72
Magnaporthe_grisea39940662
42
Neurospora_crassa40882191
Yarrowia_lipolytica50548571
38
AO2_Aspergillus_oryzae83770757


0.2
Yarrowia_lipolytica50554181
99
Candida_albicans68464685
99
AO2_Aspergillus_oryzae83772647
Aspergillus_nidulans67528025
75
62
AO1_Aspergillus_oryzae83766832
65
Aspergillus_fumigatus66848034
92
Magnaporthe_grisea39973873
78
Neurospora_crassa40804618
38
Gibberella_zeae42551031
Yarrowia_lipolytica50553306
Candida_albicans68484006
77
100
Debaryomyces_hansenii49657073
Kluyveromyces_lactis49645017
54
27
Candida_glabrata49528058
95
Candida_glabrata49524790
36
Saccharomyces_cerevisiae4824
27
100
Saccharomyces_cerevisiae1314127
25
Kluyveromyces_lactis49641039
Schizosaccharomyces_pombe1204220
Ustilago_maydis71024441

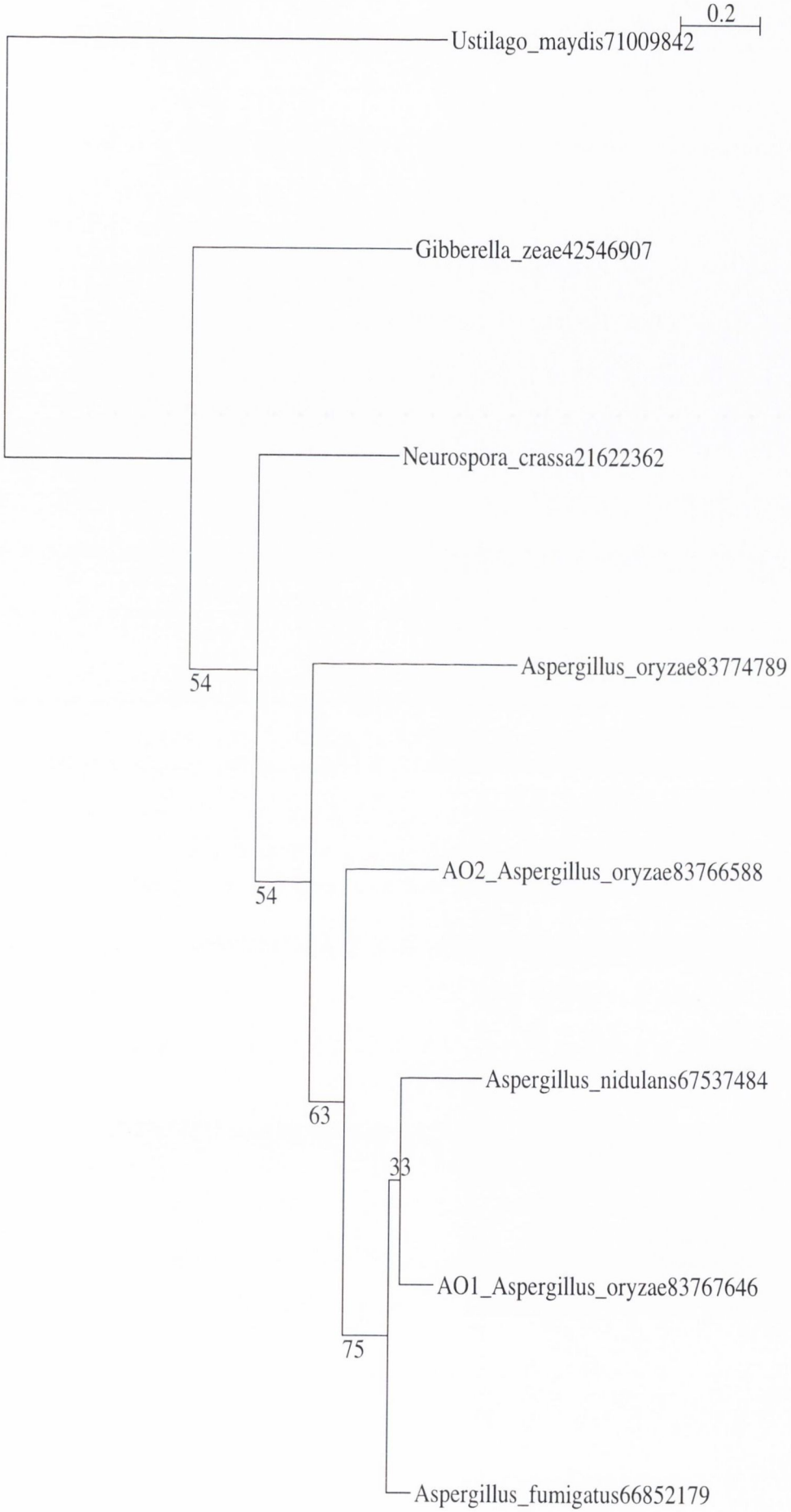

0.2
Ustilago_maydis71019279
AO2_Aspergillus_oryzae83770677
Aspergillus_nidulans67517993
89
AO1_Aspergillus_oryzae83773975
100
Aspergillus_fumigatus66850115
67
Gibberella_zeae42555196
27
Neurospora_crassa85086870
61
Gibberella_zeae42554788
49
Magnaporthe_grisea39974983

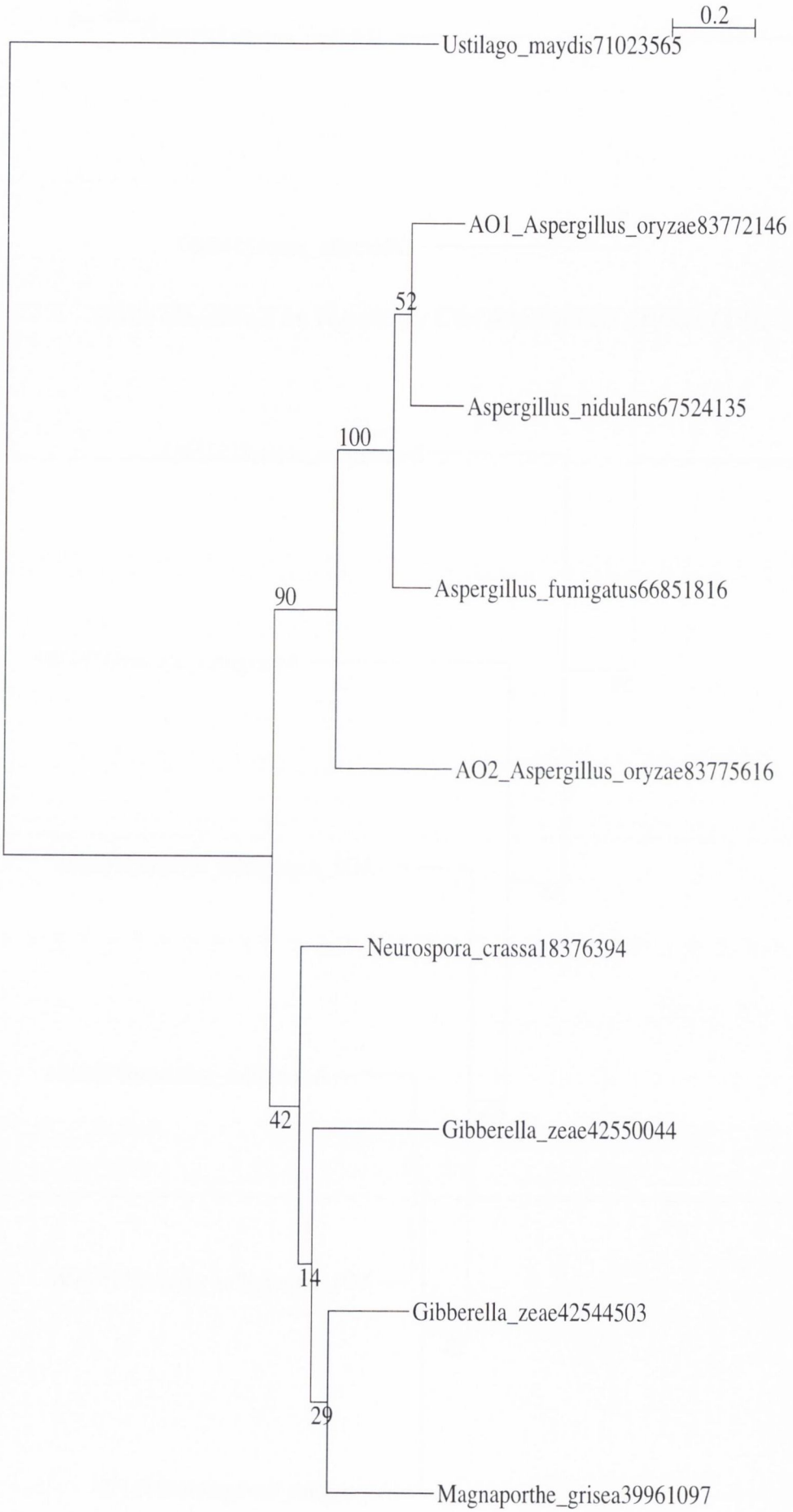

0.1
Schizosaccharomyces_pombe4455786
Schizosaccharomyces_pombe1749430
Debaryomyces_hansenii49657447
Candida_albicans68489770
Saccharomyces_cerevisiae736307
Candida_glabrata49527856
Ashbya_gossypii44985135
Kluyveromyces_lactis49644604
Yarrowia_lipolytica50546525
AO2_Aspergillus_oryzae83773061
Neurospora_crassa7800820
Magnaporthe_grisea39971771
Fusarium_oxysporum19072114
Gibberella_zeae42544553
Aspergillus_fumigatus66845549
AO1_Aspergillus_oryzae83771508
Aspergillus_nidulans67902574
Ustilago_maydis71023367
Cryptococcus_neoformans57230480
100
100
46
74
46
69
92
95
97
76
48
49
100
59
100
67

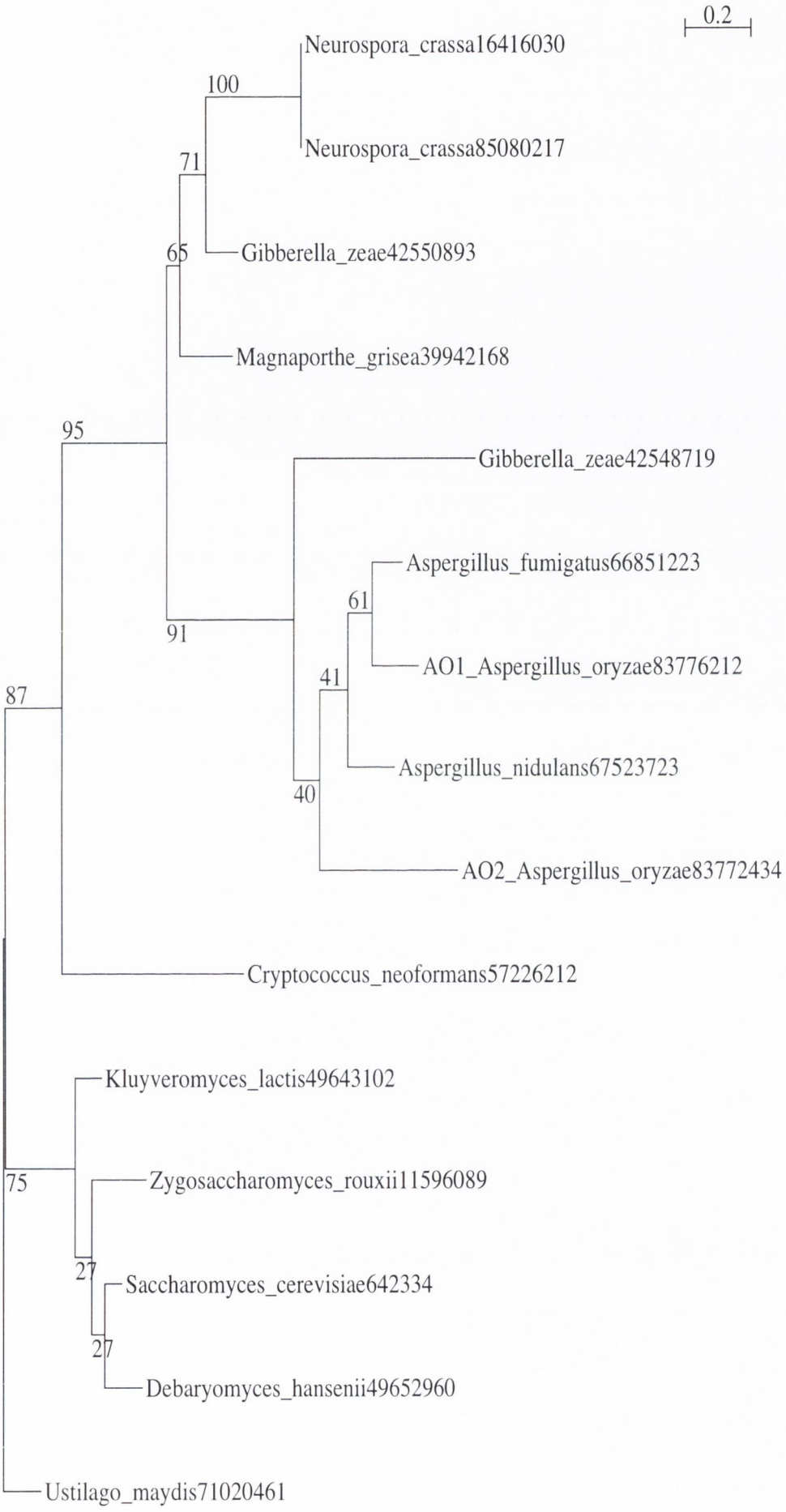

0.2
Ustilago_maydis71018647
Neurospora_crassa85103892
Aspergillus_flavus49409607
100
AO1_Aspergillus_oryzae83773082
77
Aspergillus_fumigatus66844494
94
48
Aspergillus_nidulans67541146
Gibberella_zeae42551994
77
69
Magnaporthe_grisea39949527
AO2_Aspergillus_oryzae83769592

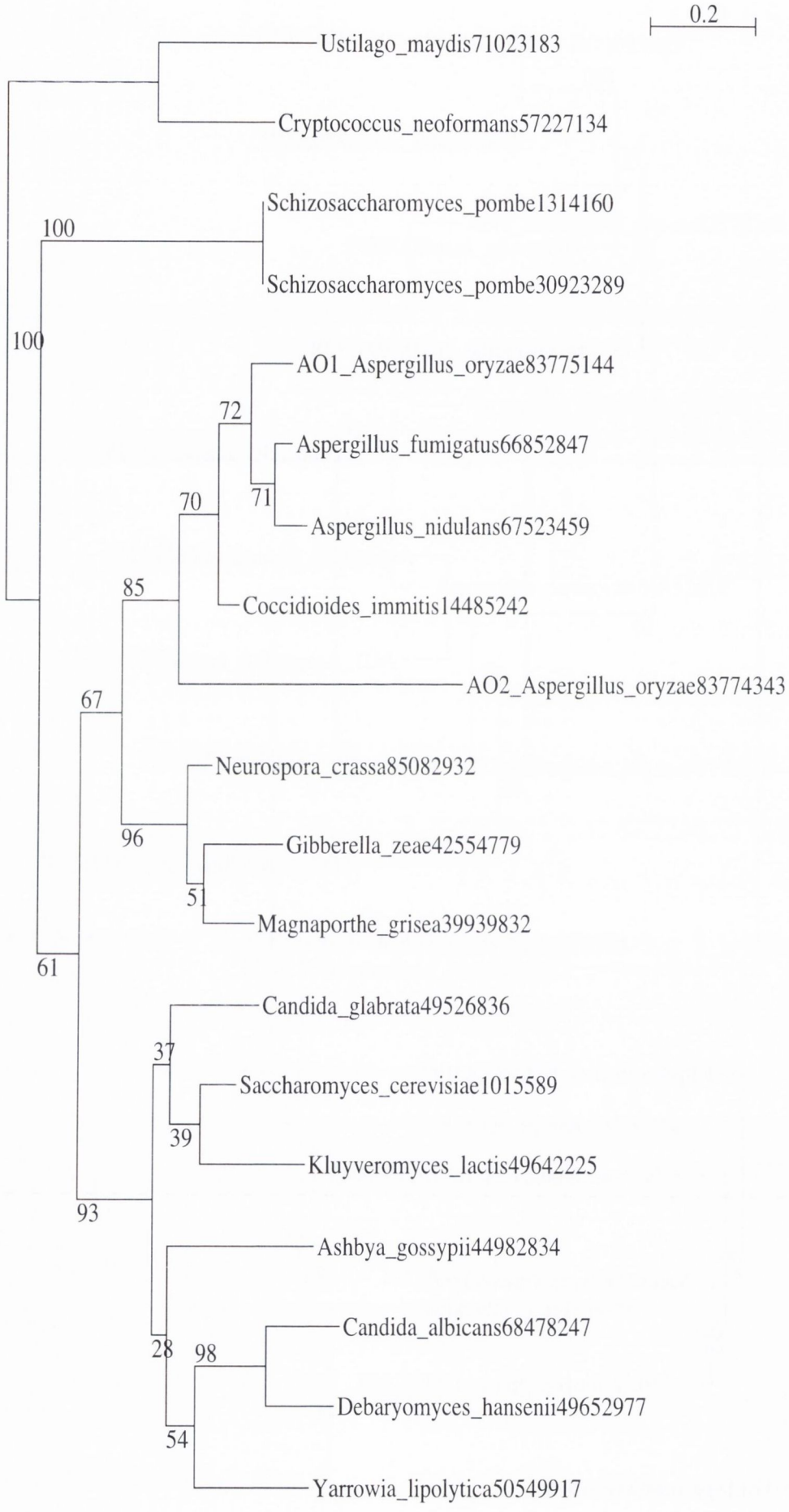

0.2
Cryptococcus_neoformans57228771
Ustilago_maydis71021797
100
AO2_Aspergillus_oryzae83776155
Neurospora_crassa85108246
99
Gibberella_zeae42551739
87
Magnaporthe_grisea39951701
65
Aspergillus_fumigatus66853652
63
Aspergillus_nidulans67900634
47
AO1_Aspergillus_oryzae83771115

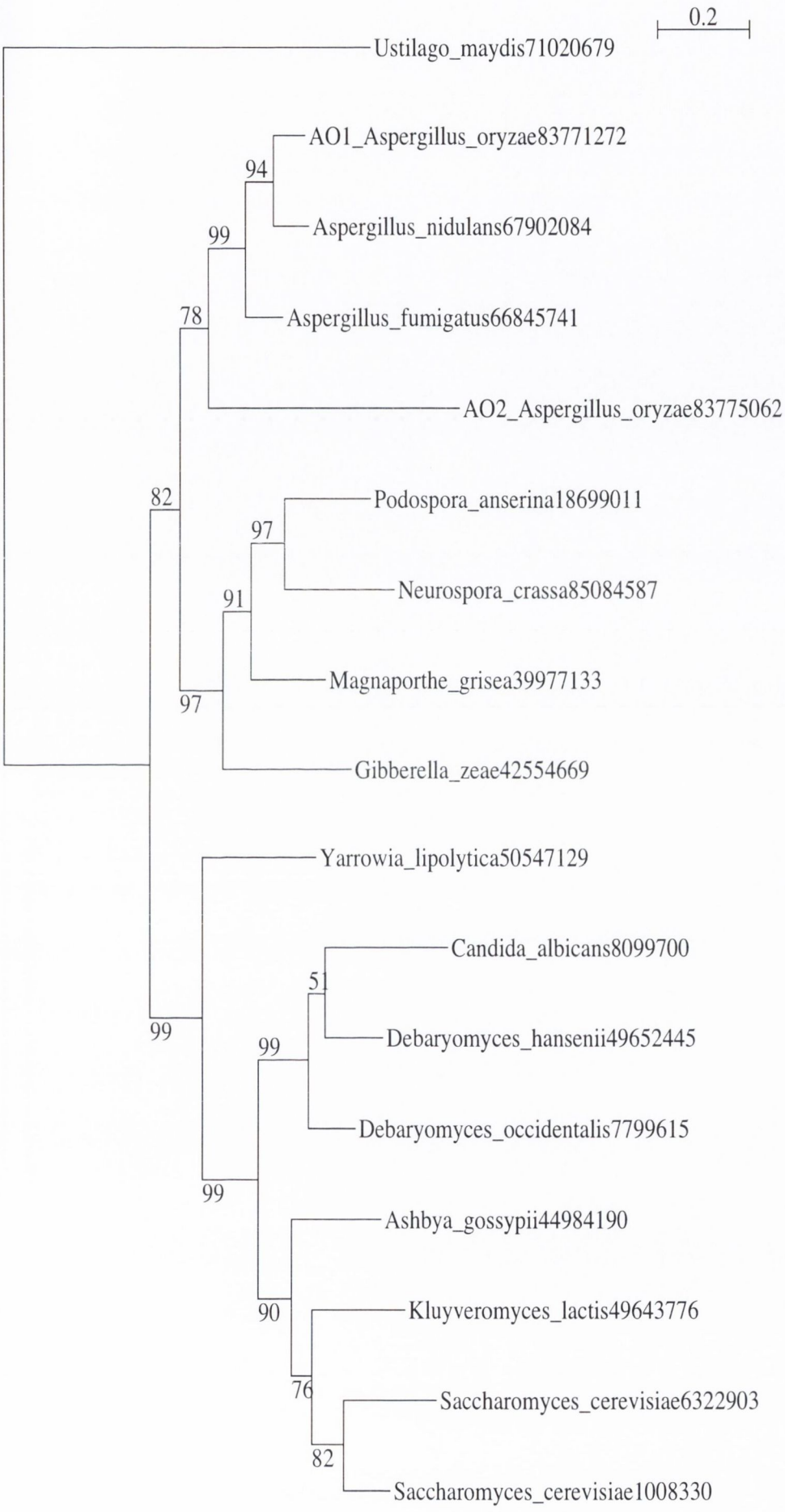
Trees classified as Topology C in *Aspergillus oryzae* (14).


0.2
Ustilago_maydis71009842
Gibberella_zeae42546907
Neurospora_crassa21622362
54
Aspergillus_oryzae83774789
54
AO2_Aspergillus_oryzae83766588
63
Aspergillus_nidulans67537484
33
AO1_Aspergillus_oryzae83767646
75
Aspergillus_fumigatus66852179

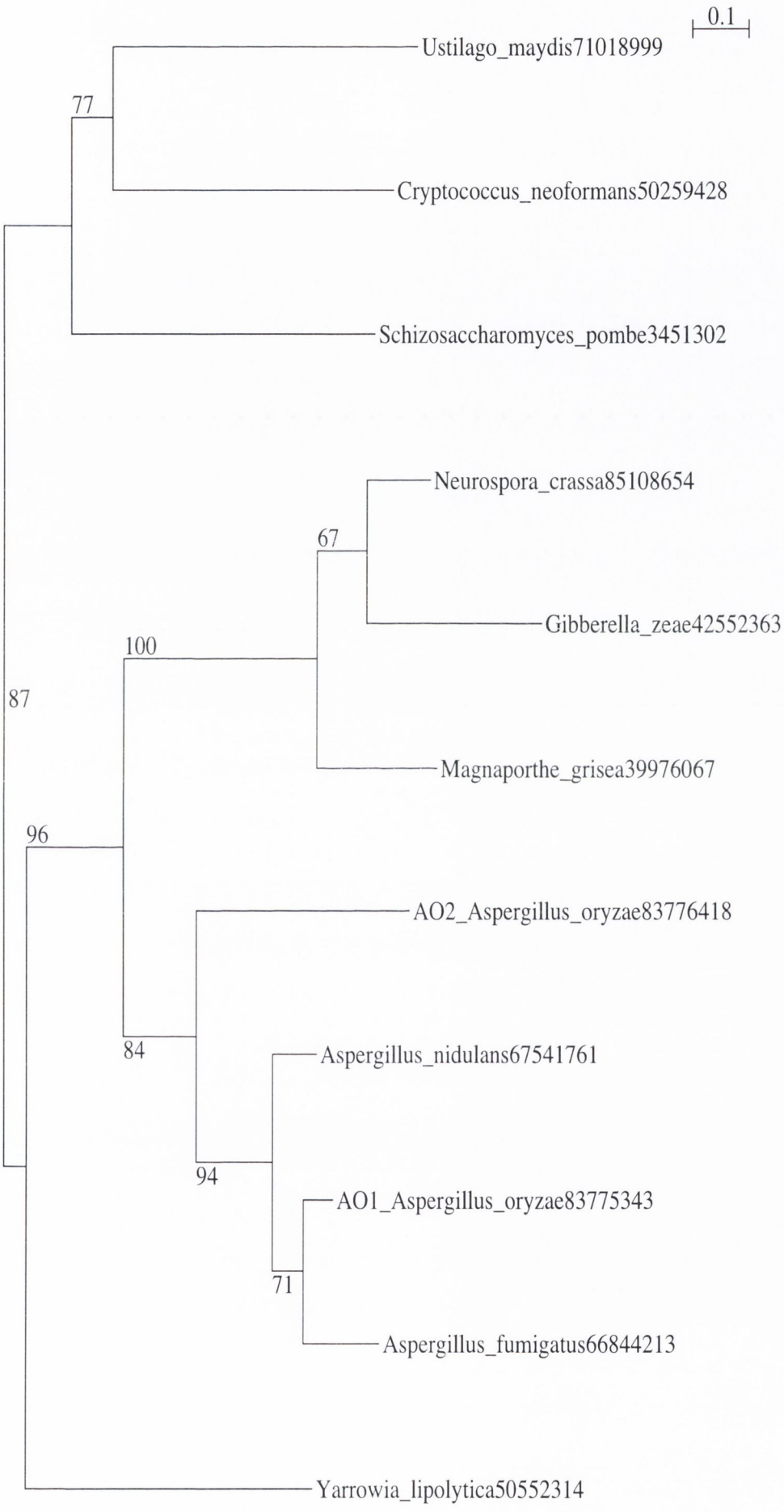

0.2
Ustilago_maydis71023565
AO1_Aspergillus_oryzae83772146
52
Aspergillus_nidulans67524135
100
Aspergillus_fumigatus66851816
90
AO2_Aspergillus_oryzae83775616
Neurospora_crassa18376394
42
Gibberella_zeae42550044
14
Gibberella_zeae42544503
29
Magnaporthe_grisea39961097

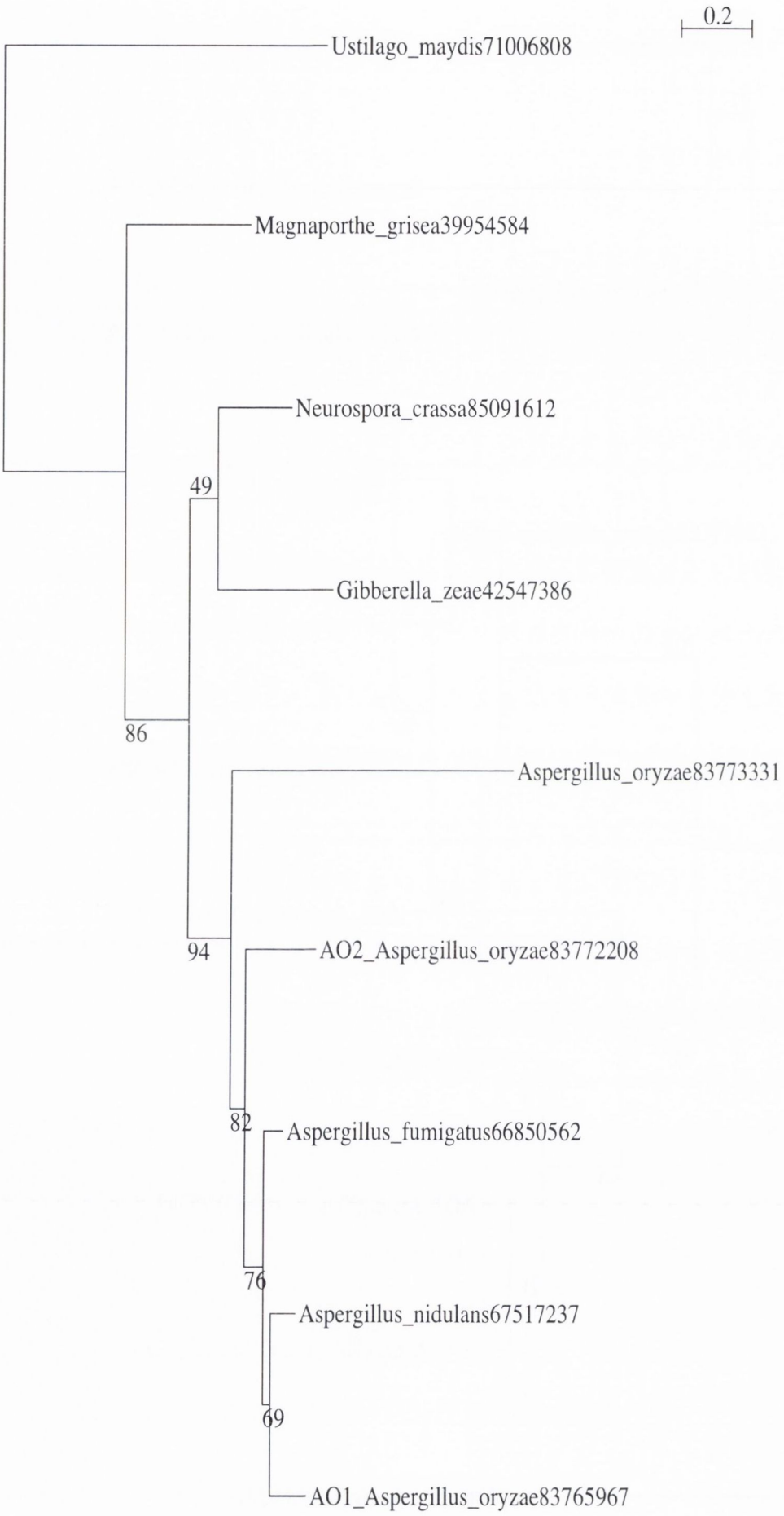

0.2
Neurospora_crassa16416030
100
Neurospora_crassa85080217
71
Gibberella_zeae42550893
65
Magnaporthe_grisea39942168
95
Gibberella_zeae42548719
Aspergillus_fumigatus66851223
61
91
AO1_Aspergillus_oryzae83776212
41
87
Aspergillus_nidulans67523723
40
AO2_Aspergillus_oryzae83772434
Cryptococcus_neoformans57226212
Kluyveromyces_lactis49643102
75
Zygosaccharomyces_rouxii11596089
27
Saccharomyces_cerevisiae642334
27
Debaryomyces_hansenii49652960
Ustilago_maydis71020461

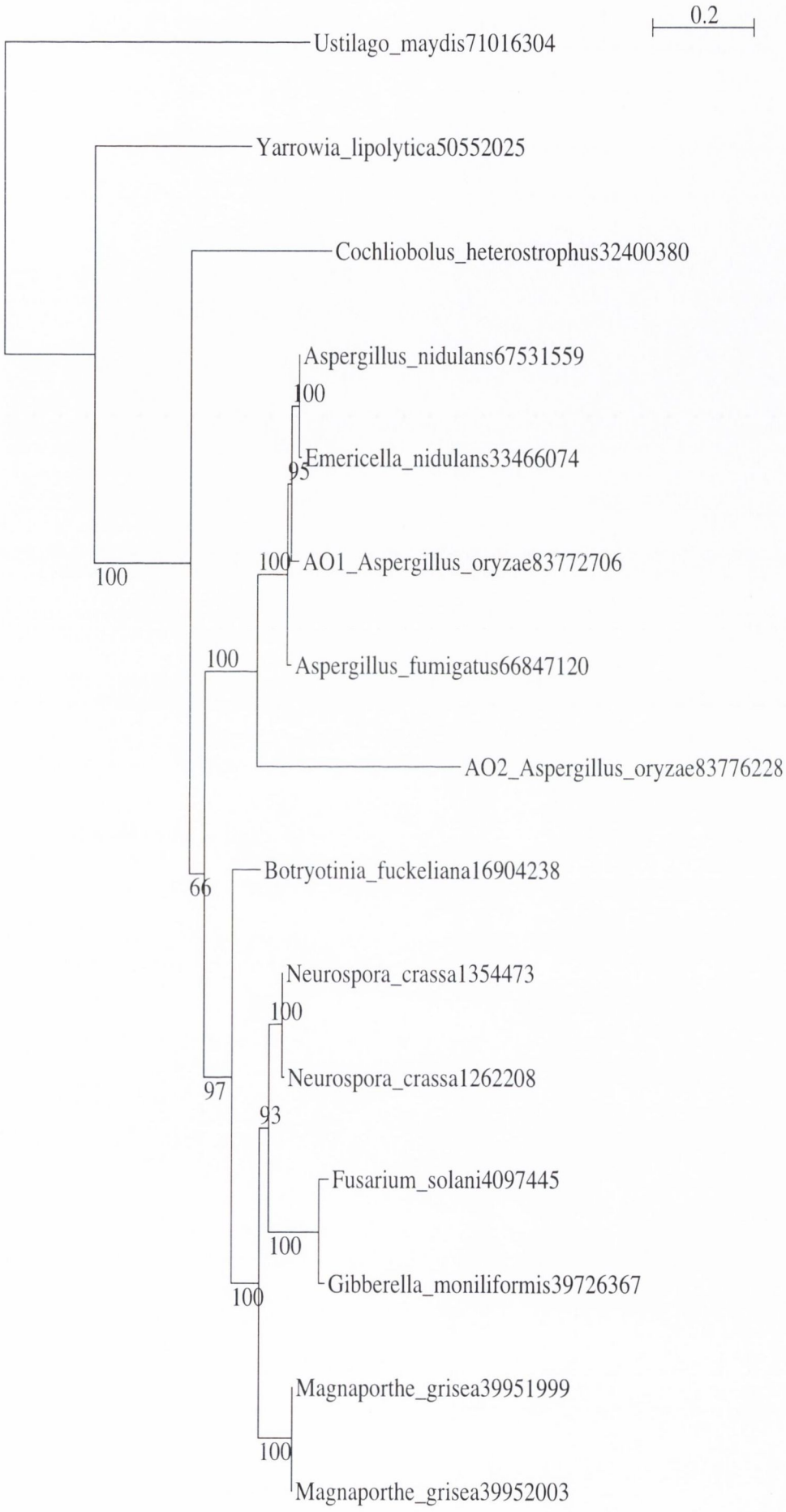

0.2
Ustilago_maydis71023183
Cryptococcus_neoformans57227134
Schizosaccharomyces_pombe1314160
Schizosaccharomyces_pombe30923289
AO1_Aspergillus_oryzae83775144
Aspergillus_fumigatus66852847
Aspergillus_nidulans67523459
Coccidioides_immitis14485242
AO2_Aspergillus_oryzae83774343
Neurospora_crassa85082932
Gibberella_zeae42554779
Magnaporthe_grisea39939832
Candida_glabrata49526836
Saccharomyces_cerevisiae1015589
Kluyveromyces_lactis49642225
Ashbya_gossypii44982834
Candida_albicans68478247
Debaryomyces_hansenii49652977
Yarrowia_lipolytica50549917
100
100
72
71
70
85
67
96
51
61
37
39
93
28
98
54

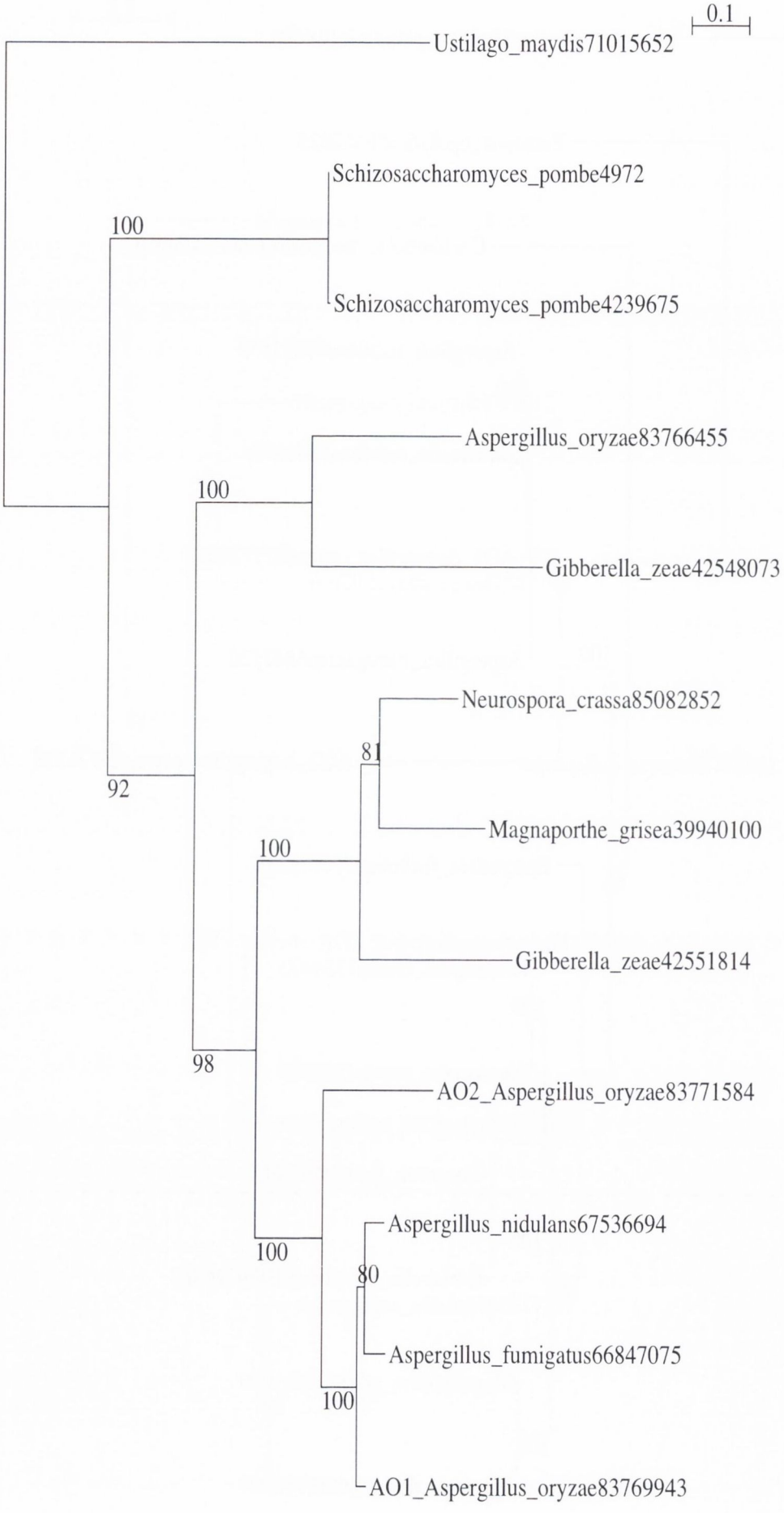

0.2
Ustilago_maydis71020679
AO1_Aspergillus_oryzae83771272
94
Aspergillus_nidulans67902084
99
Aspergillus_fumigatus66845741
78
AO2_Aspergillus_oryzae83775062
82
Podospora_anserina18699011
97
Neurospora_crassa85084587
91
Magnaporthe_grisea39977133
97
Gibberella_zeae42554669
Yarrowia_lipolytica50547129
Candida_albicans8099700
51
Debaryomyces_hansenii49652445
99
Debaryomyces_occidentalis7799615
99
99
Ashbya_gossypii44984190
90
Kluyveromyces_lactis49643776
76
Saccharomyces_cerevisiae6322903
82
Saccharomyces_cerevisiae1008330

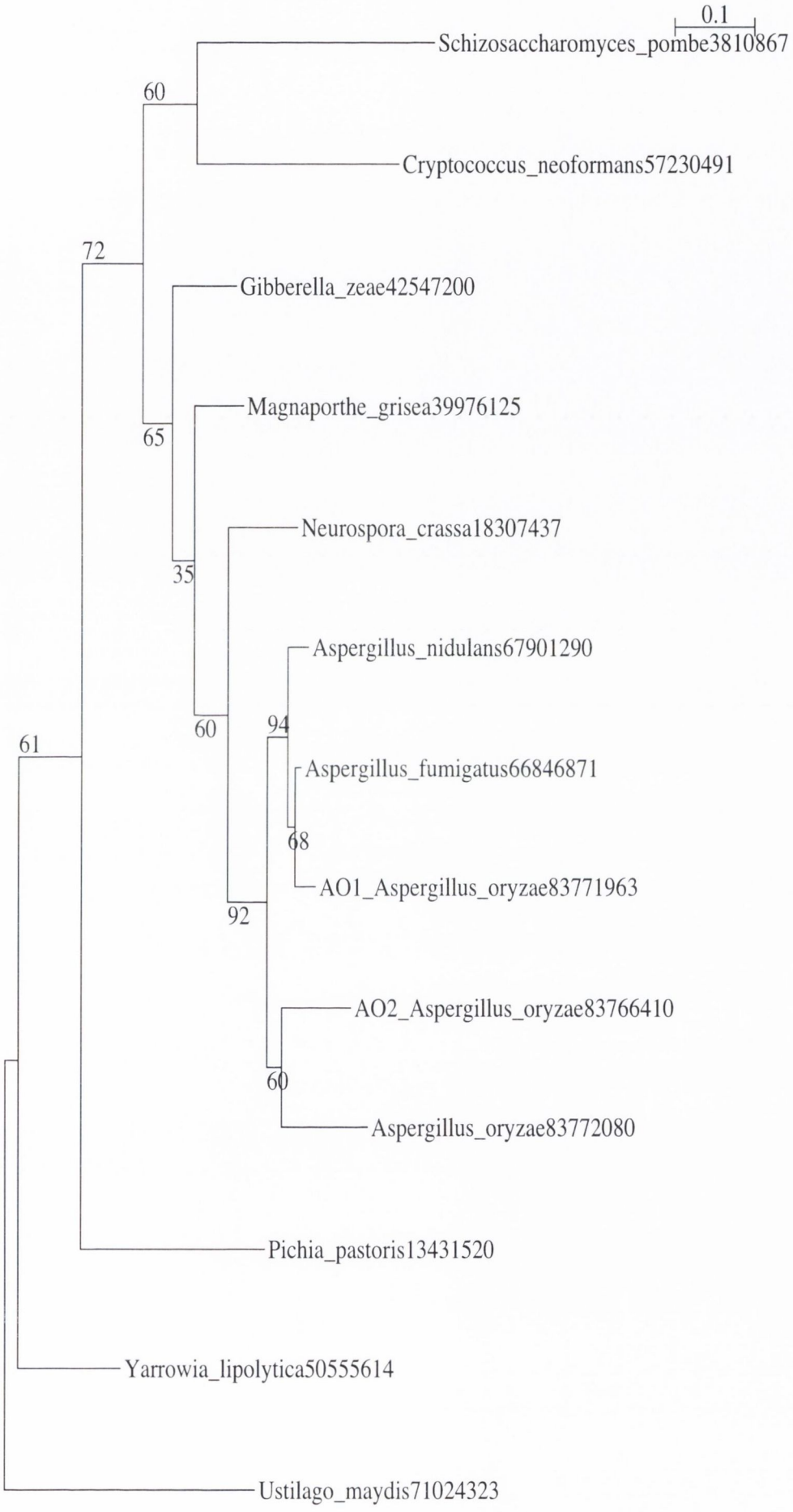
0.2
AO2_Aspergillus_oryzae83770128
100
Aspergillus_nidulans67900538
80
AO1_Aspergillus_oryzae83771083
62
Aspergillus_fumigatus28394451
100
Aspergillus_fumigatus66853682
Yarrowia_lipolytica50547955
Ustilago_maydis71003842

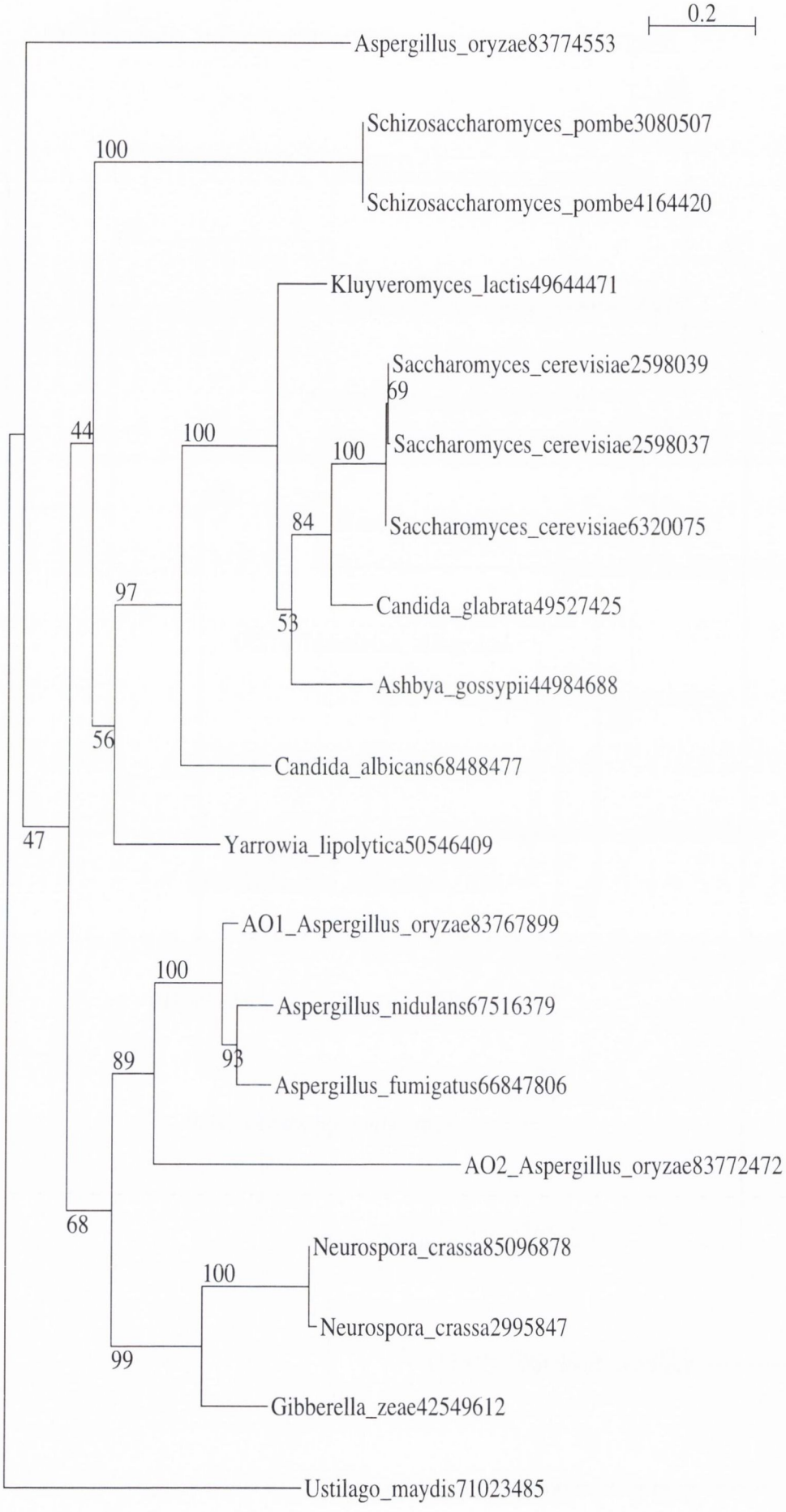

0.1
Ustilago_maydis71018999
77
Cryptococcus_neoformans50259428
Schizosaccharomyces_pombe3451302
Neurospora_crassa85108654
67
Gibberella_zeae42552363
100
87
Magnaporthe_grisea39976067
96
AO2_Aspergillus_oryzae83776418
84
Aspergillus_nidulans67541761
94
AO1_Aspergillus_oryzae83775343
71
Aspergillus_fumigatus66844213
Yarrowia_lipolytica50552314

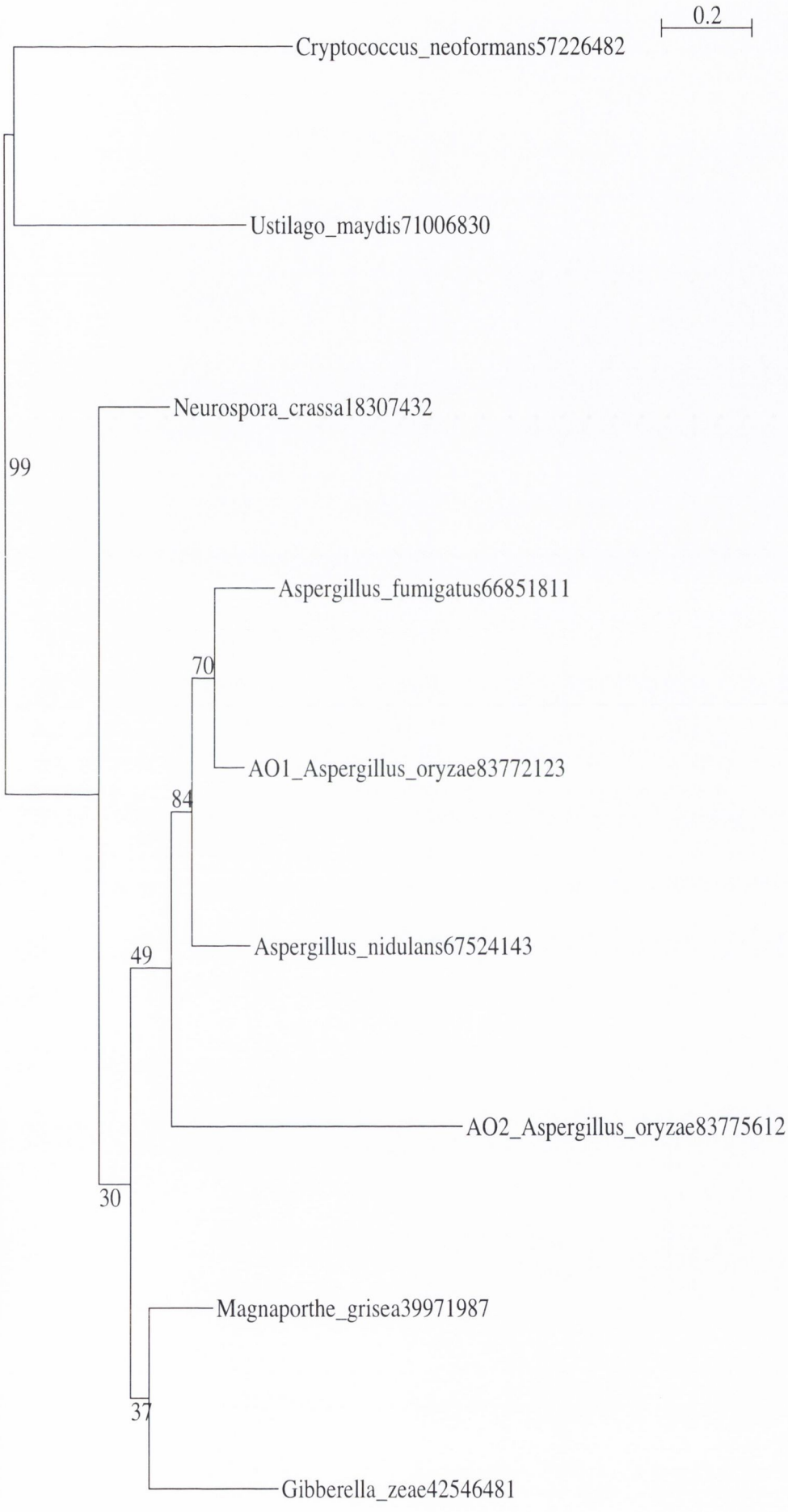

0.2
Ustilago_maydis71006808
Magnaporthe_grisea39954584
Neurospora_crassa85091612
49
Gibberella_zeae42547386
86
Aspergillus_oryzae83773331
94
AO2_Aspergillus_oryzae83772208
82
Aspergillus_fumigatus66850562
76
Aspergillus_nidulans67517237
69
AO1_Aspergillus_oryzae83765967

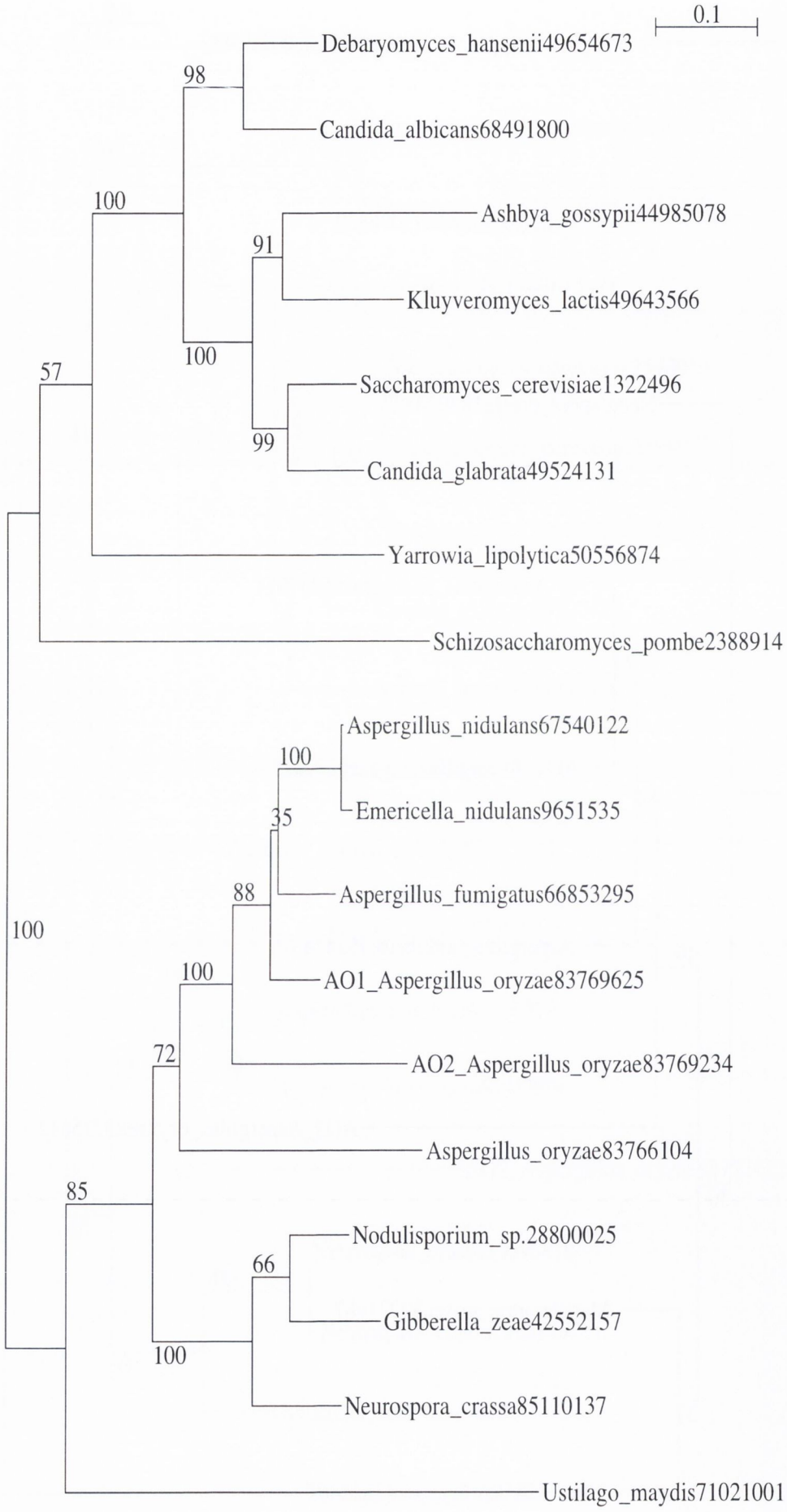

0.2
Ustilago_maydis71016304
Yarrowia_lipolytica50552025
Cochliobolus_heterostrophus32400380
Aspergillus_nidulans67531559
100
Emericella_nidulans33466074
95
100
AO1_Aspergillus_oryzae83772706
100
Aspergillus_fumigatus66847120
100
AO2_Aspergillus_oryzae83776228
Botryotinia_fuckeliana16904238
66
Neurospora_crassa1354473
100
Neurospora_crassa1262208
97
93
Fusarium_solani4097445
100
Gibberella_moniliformis39726367
100
Magnaporthe_grisea39951999
100
Magnaporthe_grisea39952003


0.1
Ustilago_maydis71015652
Schizosaccharomyces_pombe4972
100
Schizosaccharomyces_pombe4239675
Aspergillus_oryzae83766455
100
Gibberella_zeae42548073
Neurospora_crassa85082852
81
92
Magnaporthe_grisea39940100
100
Gibberella_zeae42551814
98
AO2_Aspergillus_oryzae83771584
Aspergillus_nidulans67536694
100
80
Aspergillus_fumigatus66847075
100
AO1_Aspergillus_oryzae83769943

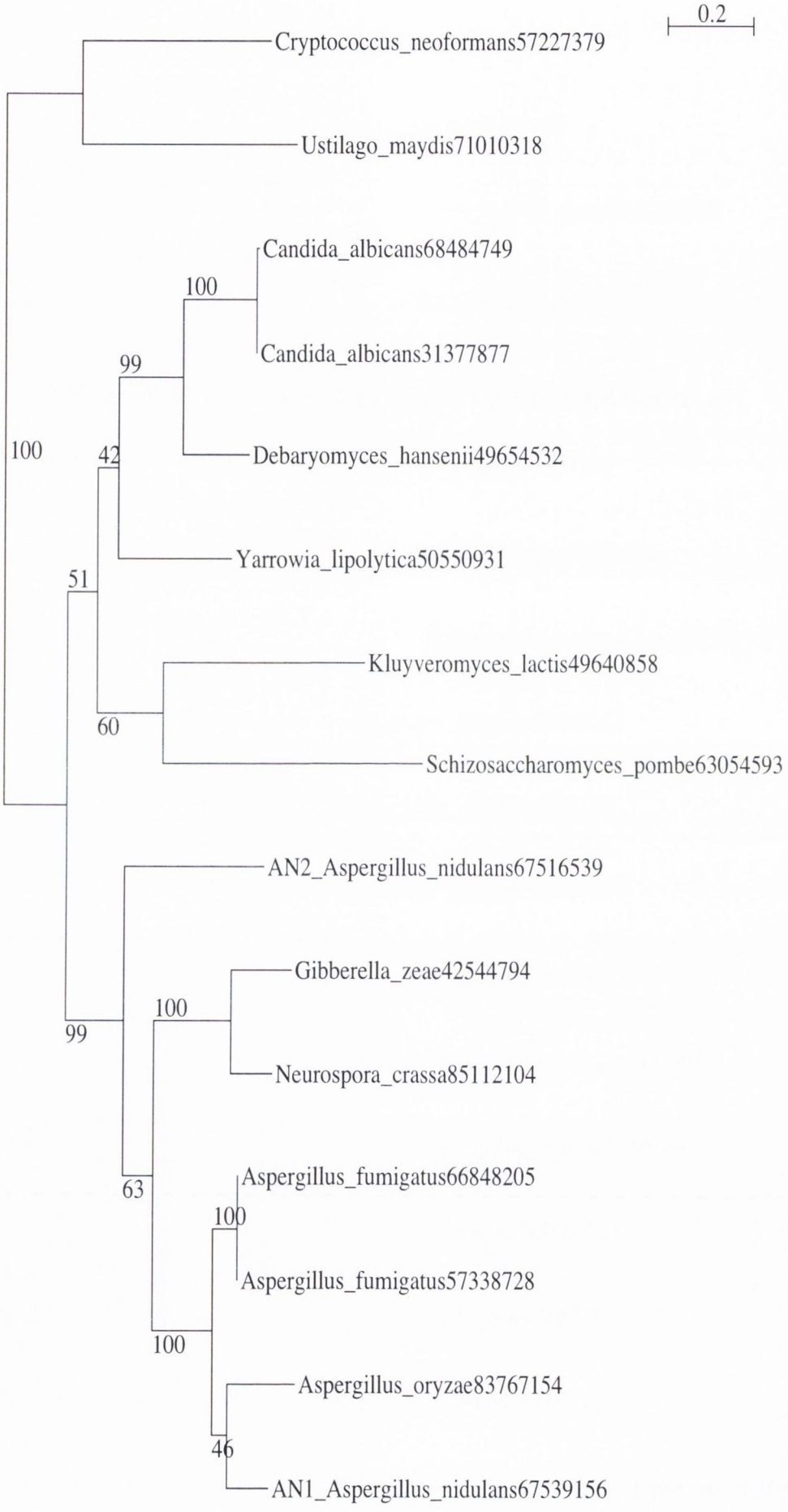

0.1
Schizosaccharomyces_pombe3810867
Cryptococcus_neoformans57230491
Gibberella_zeae42547200
Magnaporthe_grisea39976125
Neurospora_crassa18307437
Aspergillus_nidulans67901290
Aspergillus_fumigatus66846871
AO1_Aspergillus_oryzae83771963
AO2_Aspergillus_oryzae83766410
Aspergillus_oryzae83772080
Pichia_pastoris13431520
Yarrowia_lipolytica50555614
Ustilago_maydis71024323
60
72
65
35
60
94
61
68
92
60

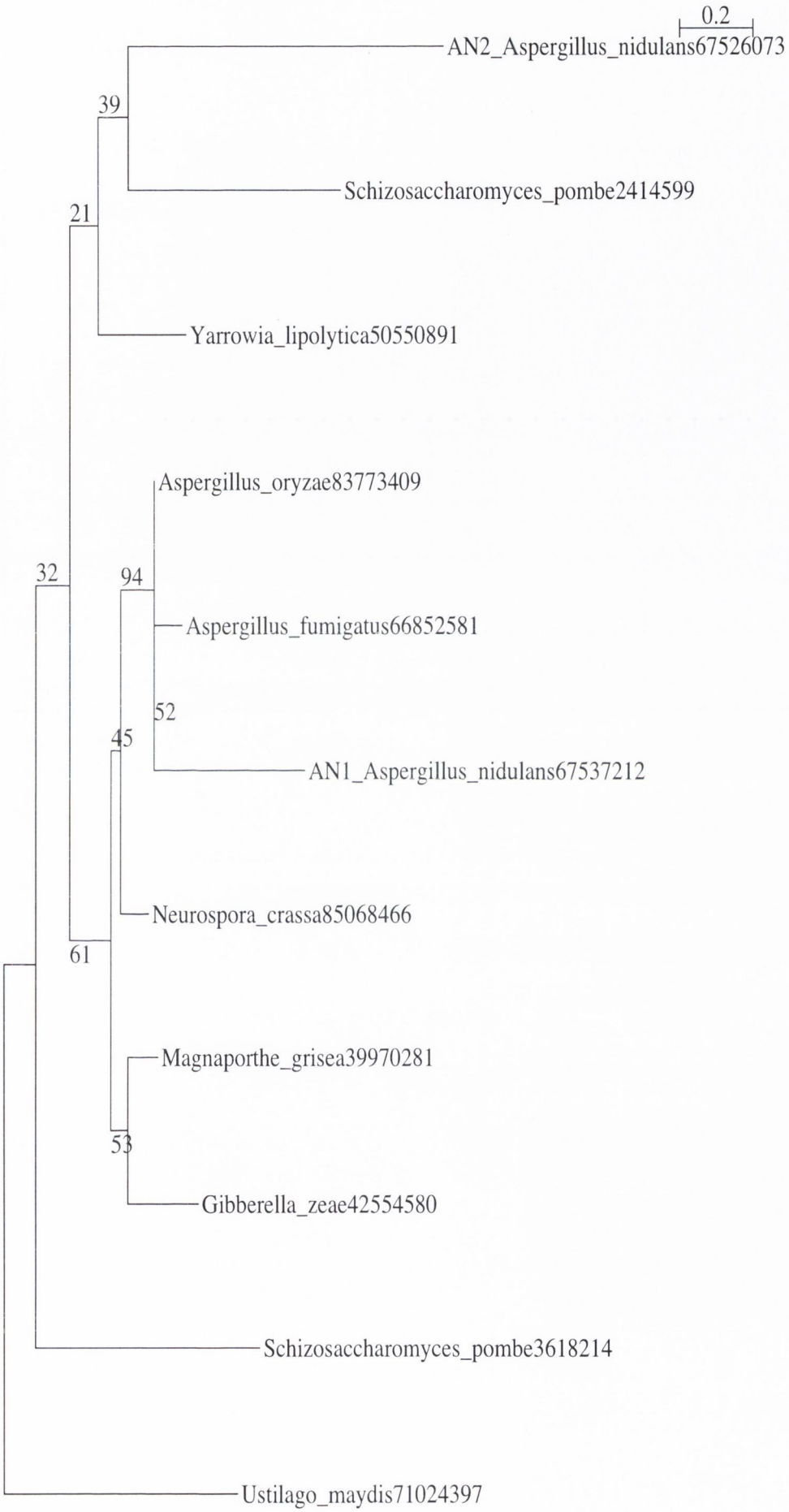

0.2
Aspergillus_oryzae83774553
Schizosaccharomyces_pombe3080507
Schizosaccharomyces_pombe4164420
Kluyveromyces_lactis49644471
Saccharomyces_cerevisiae2598039
Saccharomyces_cerevisiae2598037
Saccharomyces_cerevisiae6320075
Candida_glabrata49527425
Ashbya_gossypii44984688
Candida_albicans68488477
Yarrowia_lipolytica50546409
AO1_Aspergillus_oryzae83767899
Aspergillus_nidulans67516379
Aspergillus_fumigatus66847806
AO2_Aspergillus_oryzae83772472
Neurospora_crassa85096878
Neurospora_crassa2995847
Gibberella_zeae42549612
Ustilago_maydis71023485
100
69
44
100
100
84
97
53
56
47
100
93
89
68
100
99

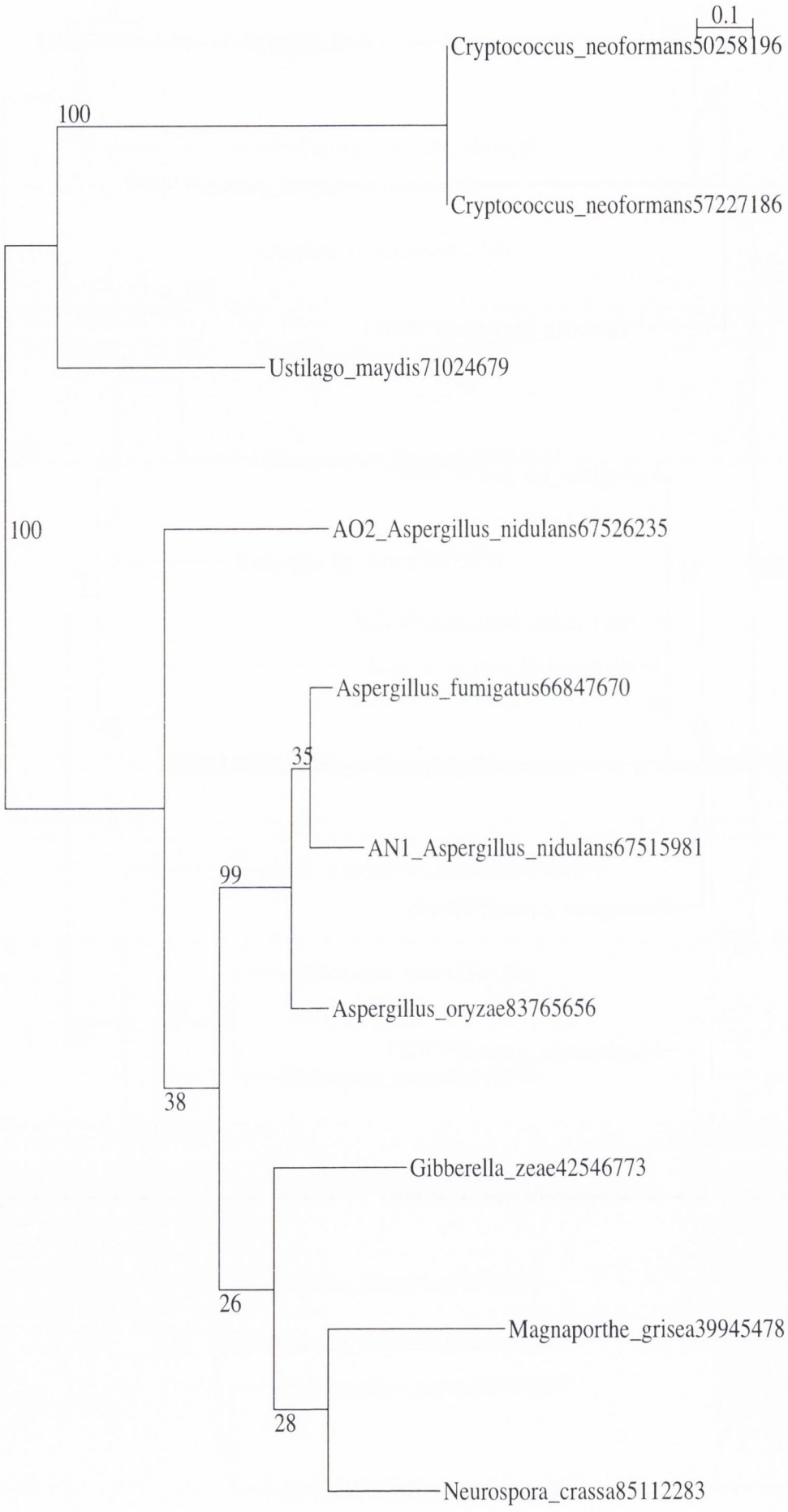

0.2
Cryptococcus_neoformans57226482
Ustilago_maydis71006830
Neurospora_crassa18307432
99
Aspergillus_fumigatus66851811
70
AO1_Aspergillus_oryzae83772123
84
Aspergillus_nidulans67524143
49
AO2_Aspergillus_oryzae83775612
30
Magnaporthe_grisea39971987
37
Gibberella_zeae42546481

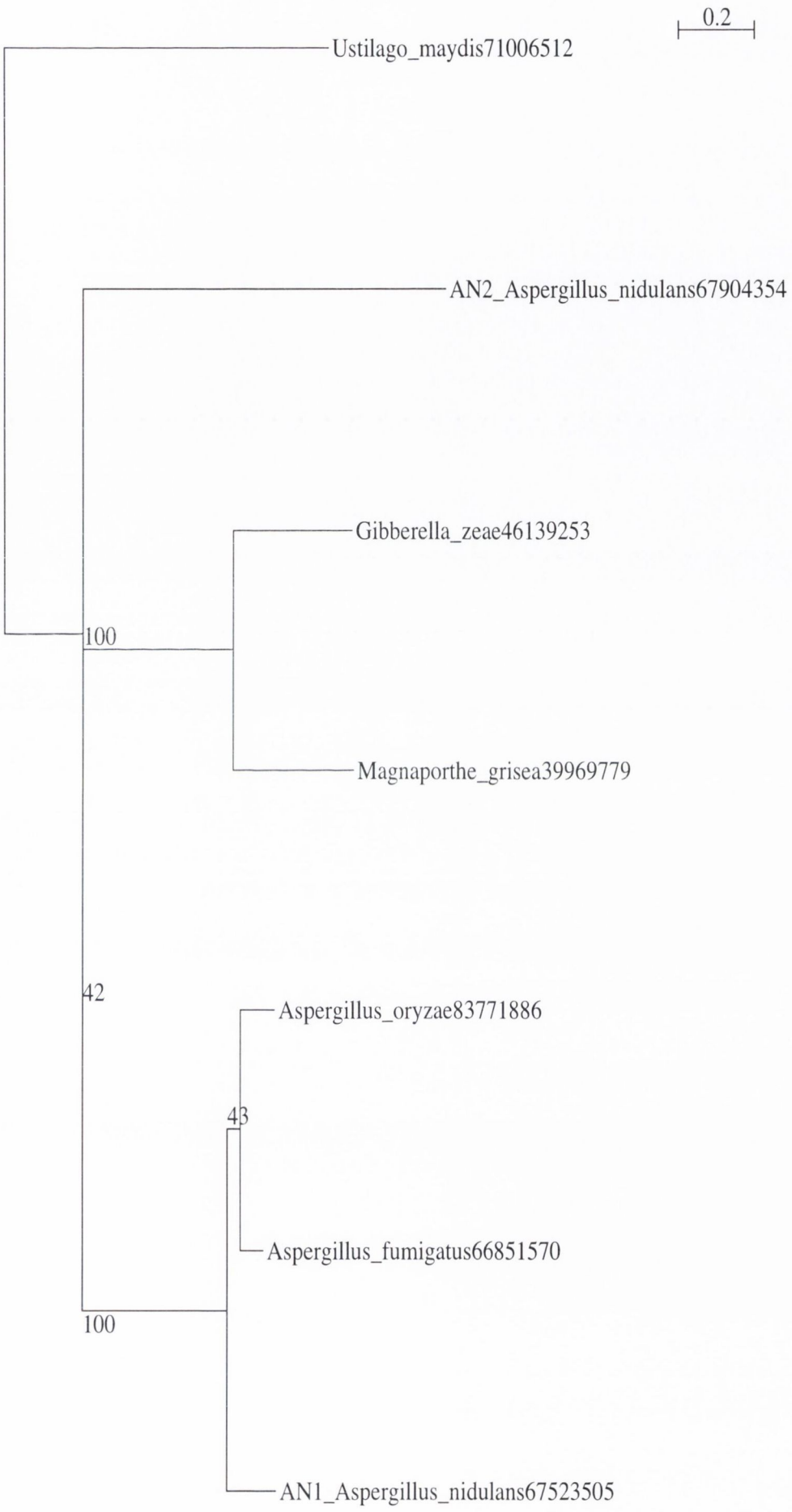

0.1
Debaryomyces_hansenii49654673
98
Candida_albicans68491800
100
Ashbya_gossypii44985078
91
Kluyveromyces_lactis49643566
100
57
Saccharomyces_cerevisiae1322496
99
Candida_glabrata49524131
Yarrowia_lipolytica50556874
Schizosaccharomyces_pombe2388914
Aspergillus_nidulans67540122
100
Emericella_nidulans9651535
35
88
Aspergillus_fumigatus66853295
100
AO1_Aspergillus_oryzae83769625
72
AO2_Aspergillus_oryzae83769234
Aspergillus_oryzae83766104
85
Nodulisporium_sp.28800025
66
Gibberella_zeae42552157
100
Neurospora_crassa85110137
Ustilago_maydis71021001

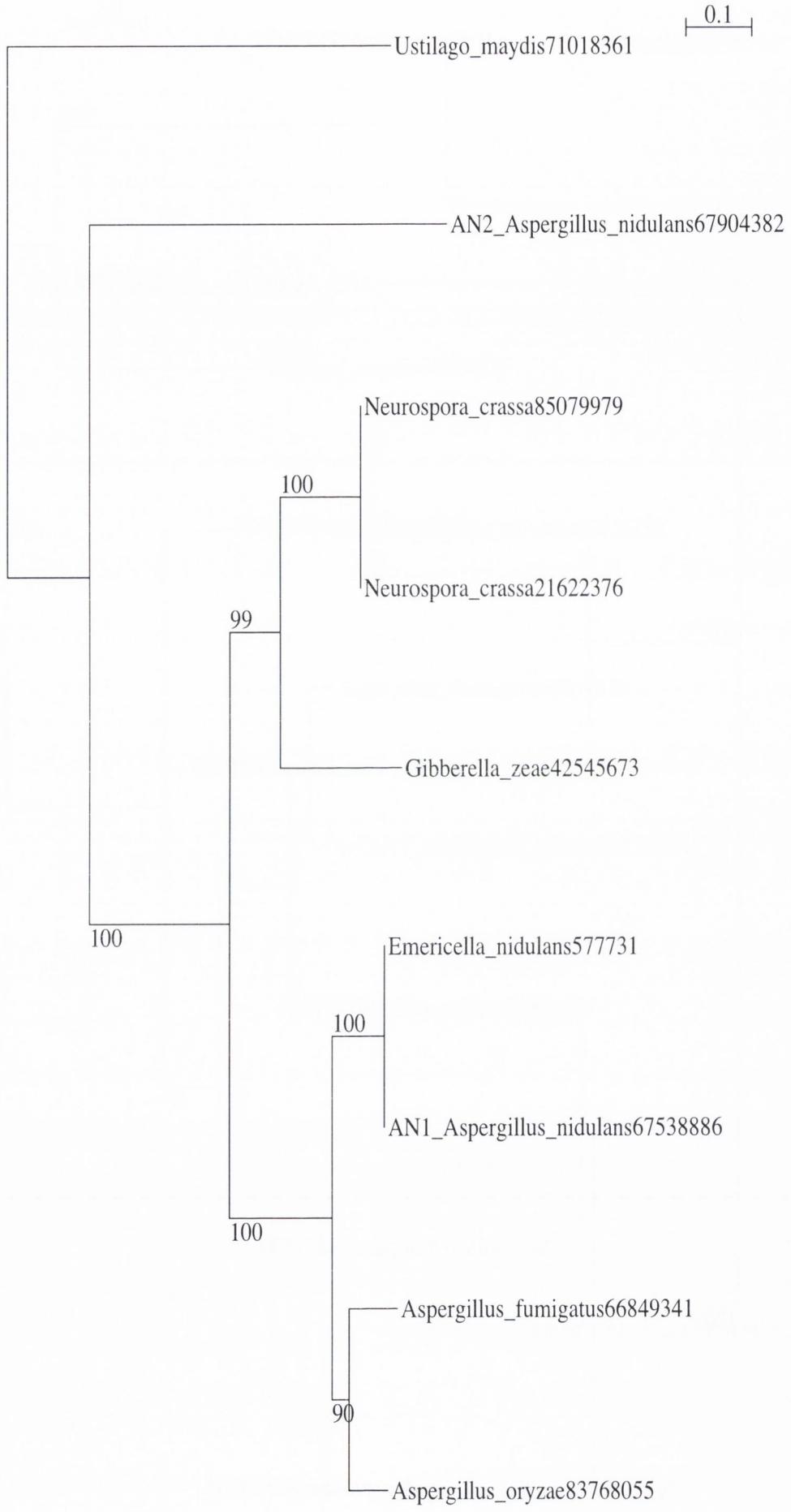
Trees classified as Topology B in *Aspergillus nidulans* (9).

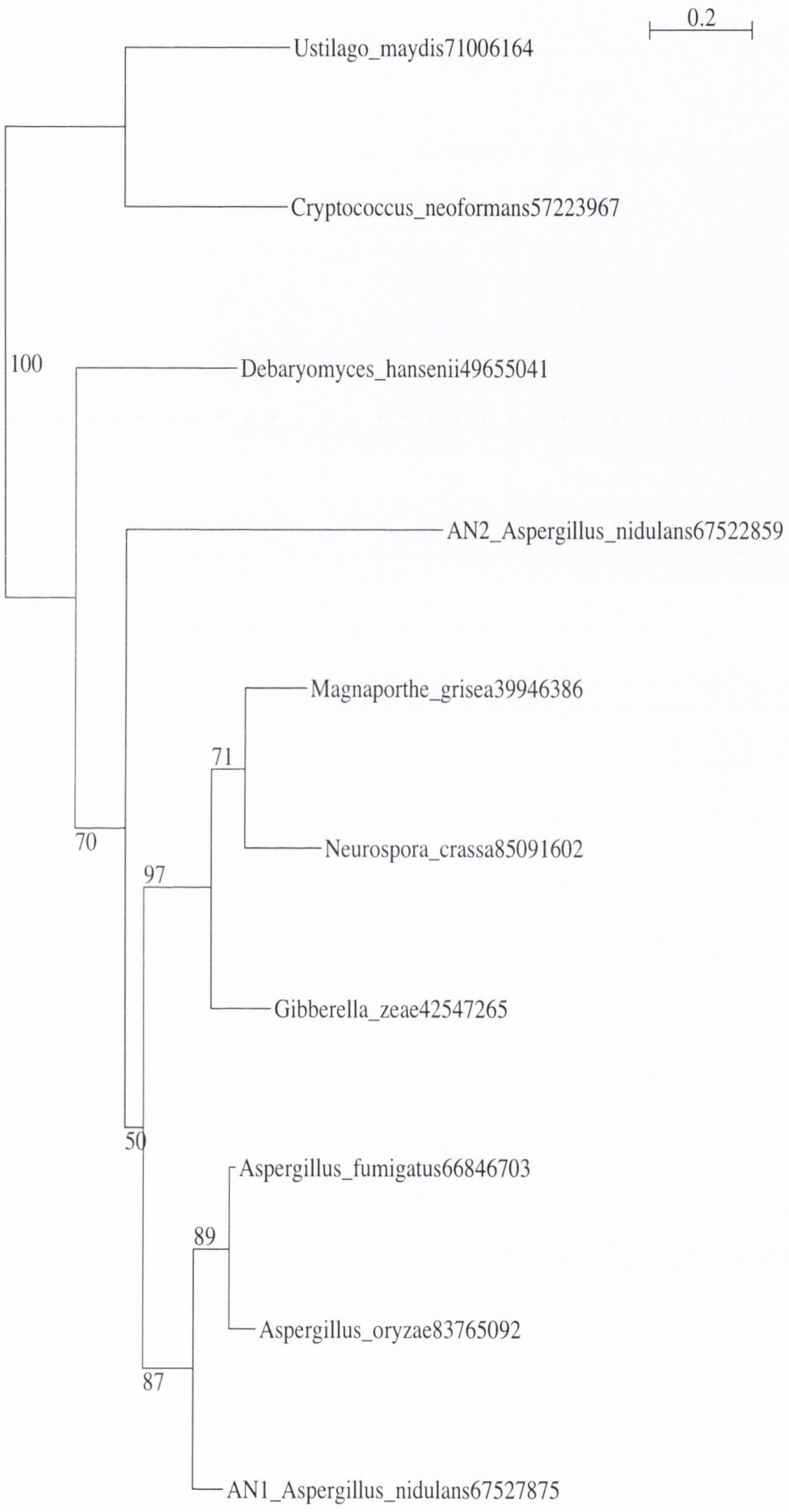

0.2
Cryptococcus_neoformans57227379
Ustilago_maydis71010318
Candida_albicans68484749
Candida_albicans31377877
Debaryomyces_hansenii49654532
Yarrowia_lipolytica50550931
Kluyveromyces_lactis49640858
Schizosaccharomyces_pombe63054593
AN2_Aspergillus_nidulans67516539
Gibberella_zeae42544794
Neurospora_crassa85112104
Aspergillus_fumigatus66848205
Aspergillus_fumigatus57338728
Aspergillus_oryzae83767154
AN1_Aspergillus_nidulans67539156
100
99
42
100
51
60
99
100
63
100
100
46

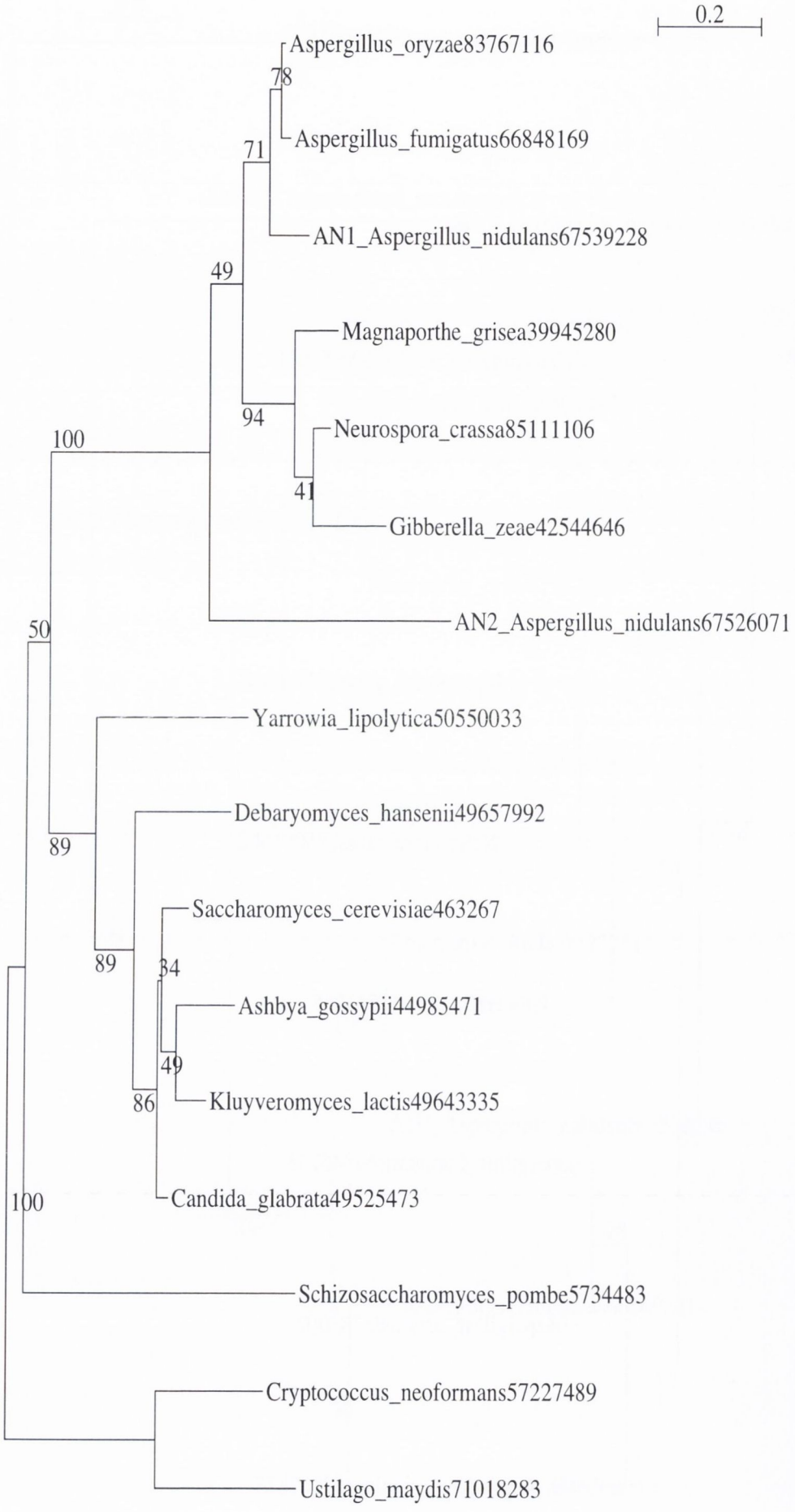

0.2
AN2_Aspergillus_nidulans67526073
39
Schizosaccharomyces_pombe2414599
21
Yarrowia_lipolytica50550891
Aspergillus_oryzae83773409
32
94
Aspergillus_fumigatus66852581
52
45
AN1_Aspergillus_nidulans67537212
Neurospora_crassa85068466
61
Magnaporthe_grisea39970281
53
Gibberella_zeae42554580
Schizosaccharomyces_pombe3618214
Ustilago_maydis71024397

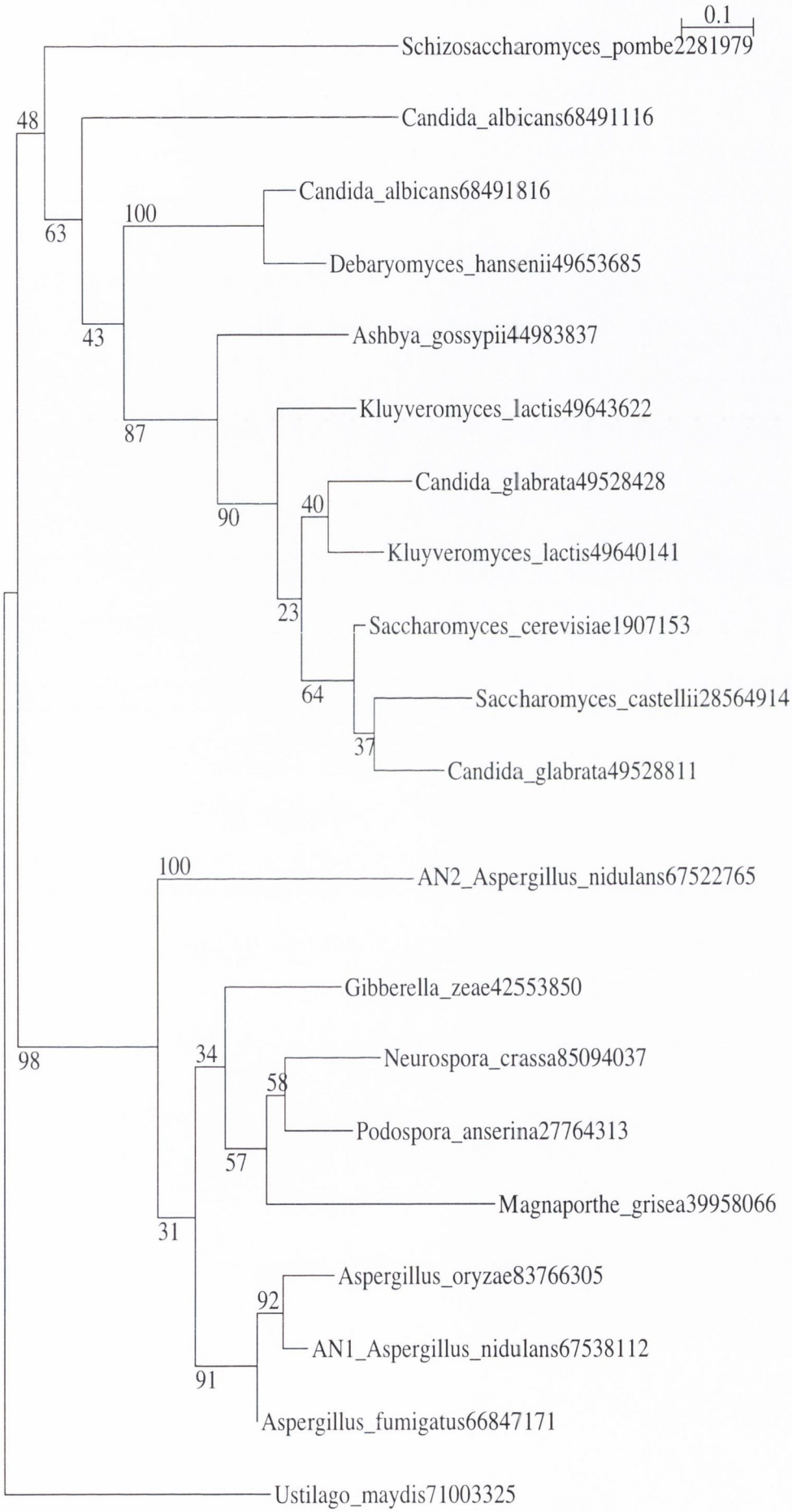

0.1
Cryptococcus_neoformans50258196
100
Cryptococcus_neoformans57227186
Ustilago_maydis71024679
100
AO2_Aspergillus_nidulans67526235
Aspergillus_fumigatus66847670
35
AN1_Aspergillus_nidulans67515981
99
Aspergillus_oryzae83765656
38
Gibberella_zeae42546773
26
Magnaporthe_grisea39945478
28
Neurospora_crassa85112283

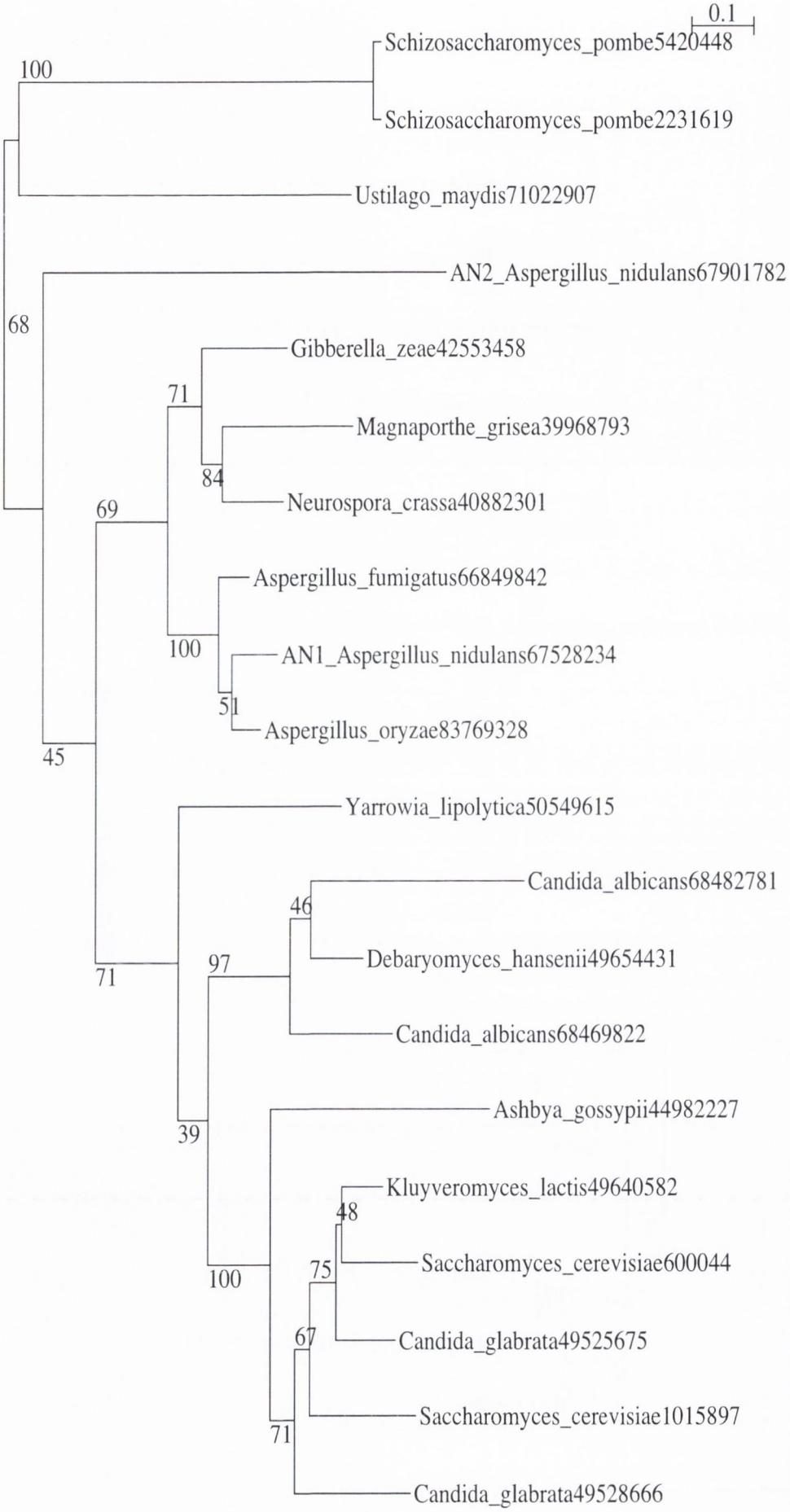

0.2
Ustilago_maydis71006512
AN2_Aspergillus_nidulans67904354
Gibberella_zeae46139253
100
Magnaporthe_grisea39969779
42
Aspergillus_oryzae83771886
43
Aspergillus_fumigatus66851570
100
AN1_Aspergillus_nidulans67523505


0.1
Ustilago_maydis71018361
AN2_Aspergillus_nidulans67904382
Neurospora_crassa85079979
100
Neurospora_crassa21622376
99
Gibberella_zeae42545673
100
Emericella_nidulans577731
100
AN1_Aspergillus_nidulans67538886
100
Aspergillus_fumigatus66849341
90
Aspergillus_oryzae83768055

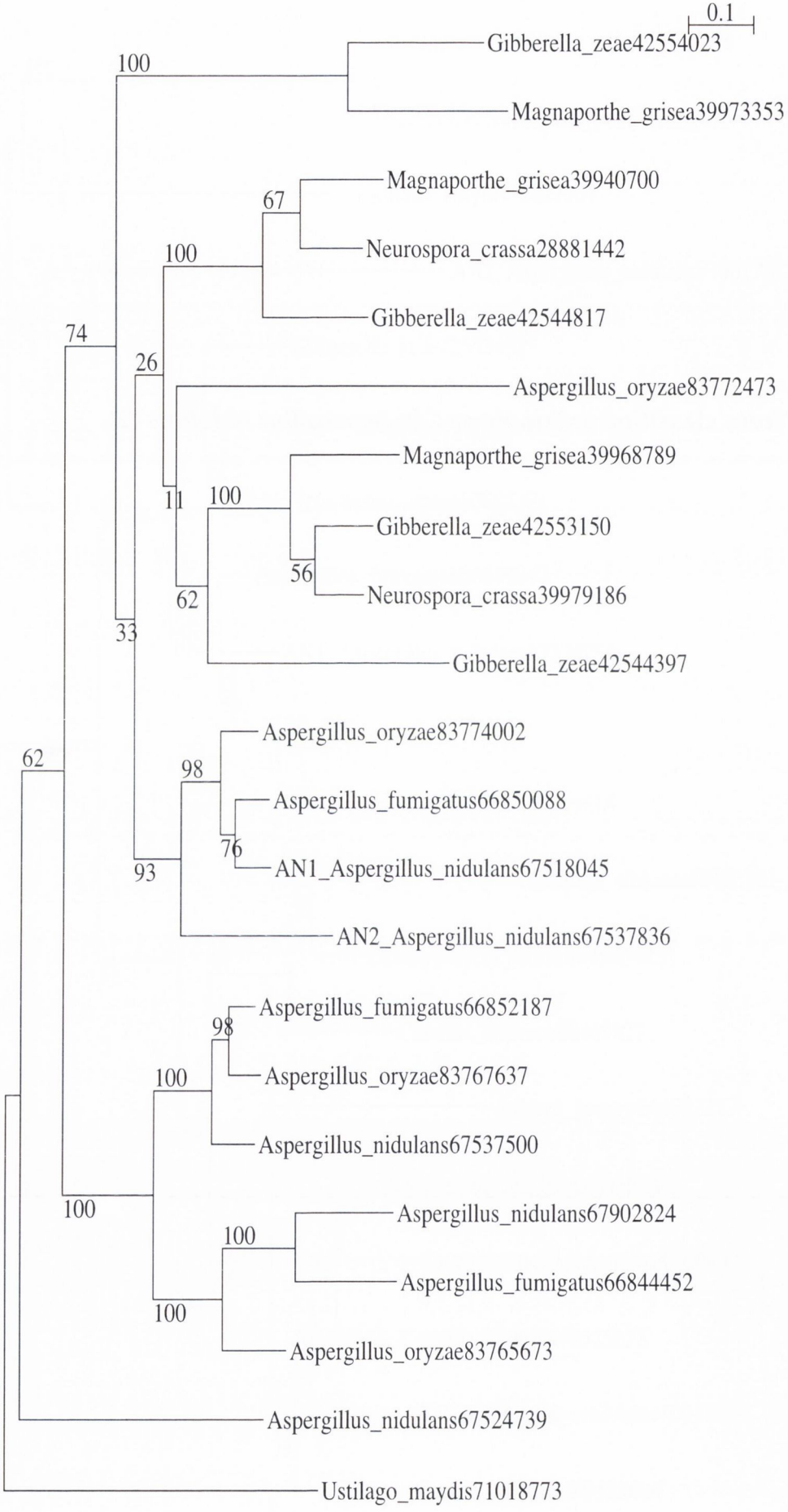

0.2
Ustilago_maydis71006164
Cryptococcus_neoformans57223967
100
Debaryomyces_hansenii49655041
AN2_Aspergillus_nidulans67522859
Magnaporthe_grisea39946386
71
Neurospora_crassa85091602
70
97
Gibberella_zeae42547265
50
Aspergillus_fumigatus66846703
89
Aspergillus_oryzae83765092
87
AN1_Aspergillus_nidulans67527875

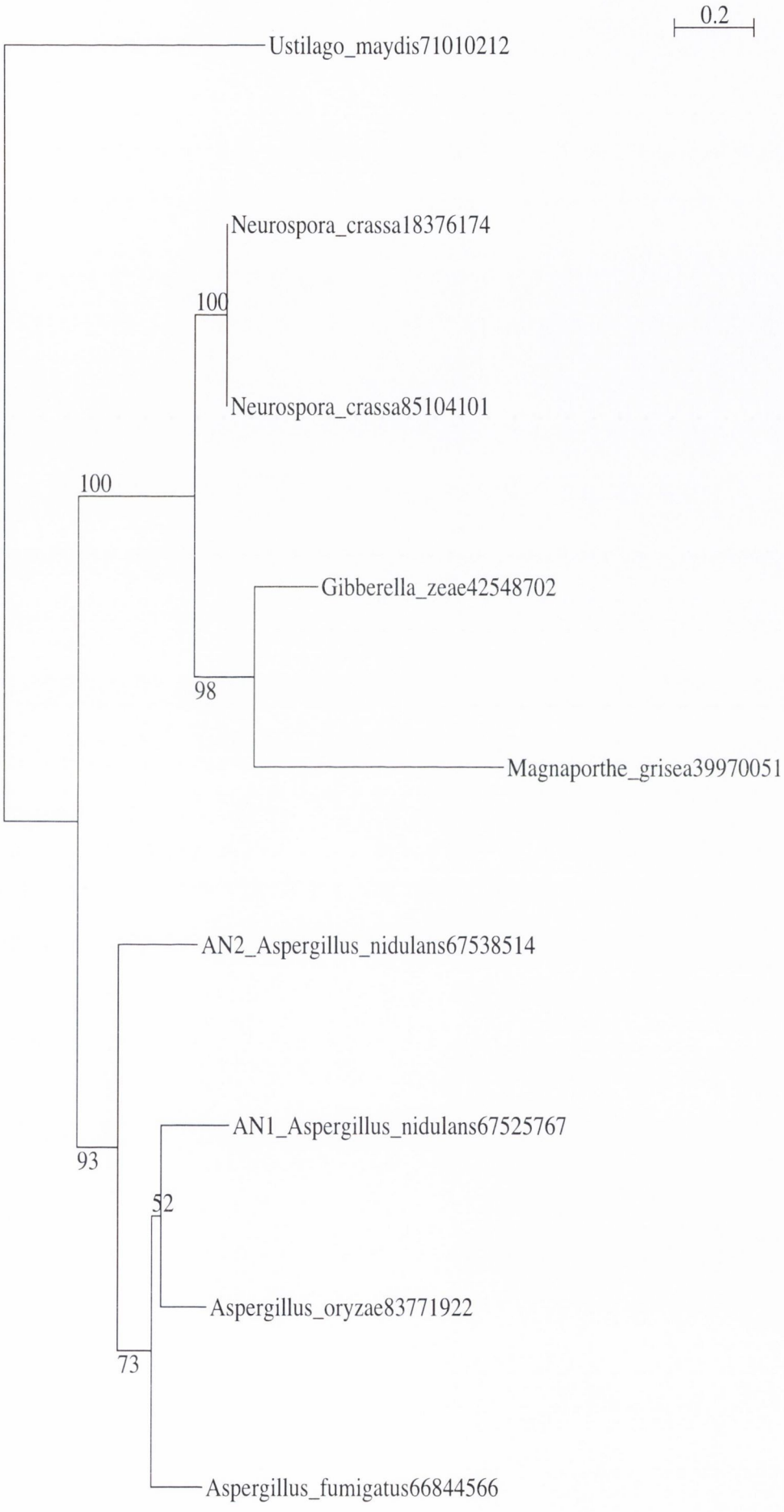

0.2
Aspergillus_oryzae83767116
78
Aspergillus_fumigatus66848169
71
AN1_Aspergillus_nidulans67539228
49
Magnaporthe_grisea39945280
94
Neurospora_crassa85111106
100
41
Gibberella_zeae42544646
AN2_Aspergillus_nidulans67526071
50
Yarrowia_lipolytica50550033
Debaryomyces_hansenii49657992
89
Saccharomyces_cerevisiae463267
89
34
Ashbya_gossypii44985471
49
86
Kluyveromyces_lactis49643335
Candida_glabrata49525473
100
Schizosaccharomyces_pombe5734483
Cryptococcus_neoformans57227489
Ustilago_maydis71018283


0.1
Schizosaccharomyces_pombe2281979
Candida_albicans68491116
Candida_albicans68491816
Debaryomyces_hansenii49653685
Ashbya_gossypii44983837
Kluyveromyces_lactis49643622
Candida_glabrata49528428
Kluyveromyces_lactis49640141
Saccharomyces_cerevisiae1907153
Saccharomyces_castellii28564914
Candida_glabrata49528811
AN2_Aspergillus_nidulans67522765
Gibberella_zeae42553850
Neurospora_crassa85094037
Podospora_anserina27764313
Magnaporthe_grisea39958066
Aspergillus_oryzae83766305
AN1_Aspergillus_nidulans67538112
Aspergillus_fumigatus66847171
Ustilago_maydis71003325
48
63
100
43
87
90
40
23
64
37
100
98
34
58
57
31
92
91

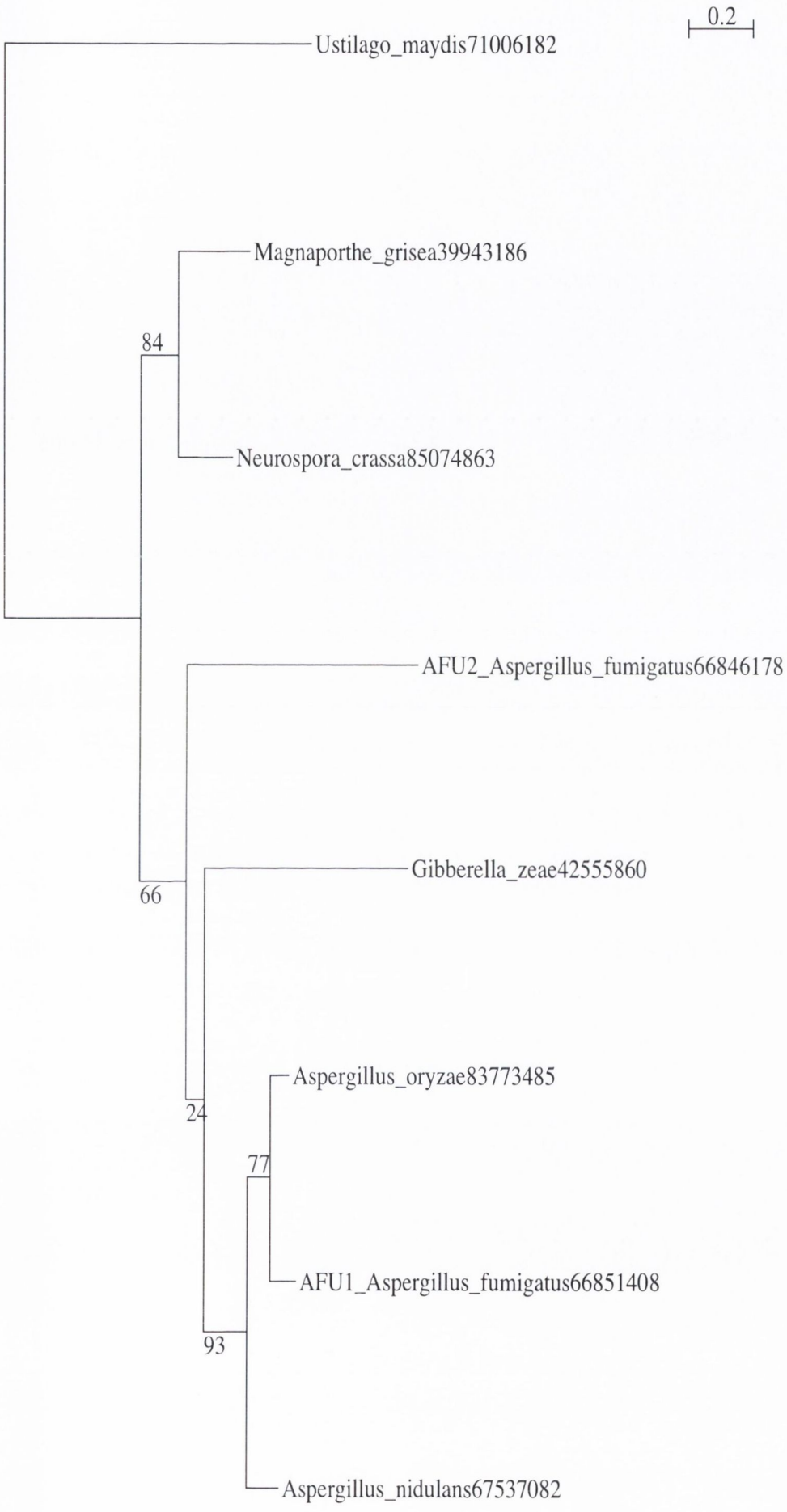

0.1
Schizosaccharomyces_pombe5420448
Schizosaccharomyces_pombe2231619
Ustilago_maydis71022907
AN2_Aspergillus_nidulans67901782
Gibberella_zeae42553458
Magnaporthe_grisea39968793
Neurospora_crassa40882301
Aspergillus_fumigatus66849842
AN1_Aspergillus_nidulans67528234
Aspergillus_oryzae83769328
Yarrowia_lipolytica50549615
Candida_albicans68482781
Debaryomyces_hansenii49654431
Candida_albicans68469822
Ashbya_gossypii44982227
Kluyveromyces_lactis49640582
Saccharomyces_cerevisiae600044
Candida_glabrata49525675
Saccharomyces_cerevisiae1015897
Candida_glabrata49528666
100
68
71
84
69
100
51
45
71
46
97
39
100
48
75
67
71

Trees classified as Topology C in *Aspergillus nidulans* (2).


0.1
Gibberella_zeae42554023
Magnaporthe_grisea39973353
Magnaporthe_grisea39940700
Neurospora_crassa28881442
Gibberella_zeae42544817
Aspergillus_oryzae83772473
Magnaporthe_grisea39968789
Gibberella_zeae42553150
Neurospora_crassa39979186
Gibberella_zeae42544397
Aspergillus_oryzae83774002
Aspergillus_fumigatus66850088
AN1_Aspergillus_nidulans67518045
AN2_Aspergillus_nidulans67537836
Aspergillus_fumigatus66852187
Aspergillus_oryzae83767637
Aspergillus_nidulans67537500
Aspergillus_nidulans67902824
Aspergillus_fumigatus66844452
Aspergillus_oryzae83765673
Aspergillus_nidulans67524739
Ustilago_maydis71018773
100
67
100
74
26
11
100
56
62
33
62
98
76
93
98
100
100
100
100

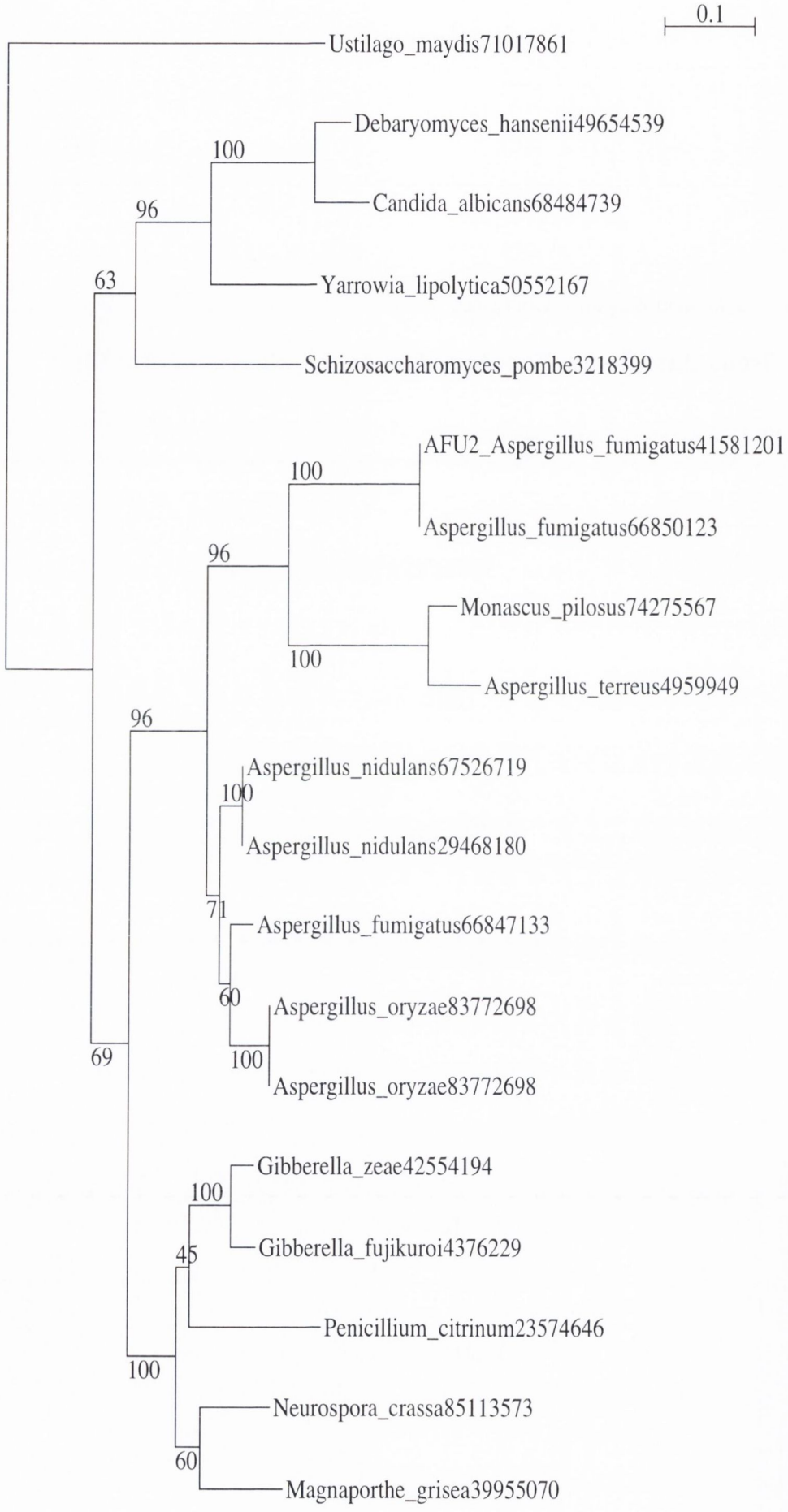

0.2
Ustilago_maydis71010212
Neurospora_crassa18376174
100
Neurospora_crassa85104101
100
Gibberella_zeae42548702
98
Magnaporthe_grisea39970051
AN2_Aspergillus_nidulans67538514
AN1_Aspergillus_nidulans67525767
93
52
Aspergillus_oryzae83771922
73
Aspergillus_fumigatus66844566

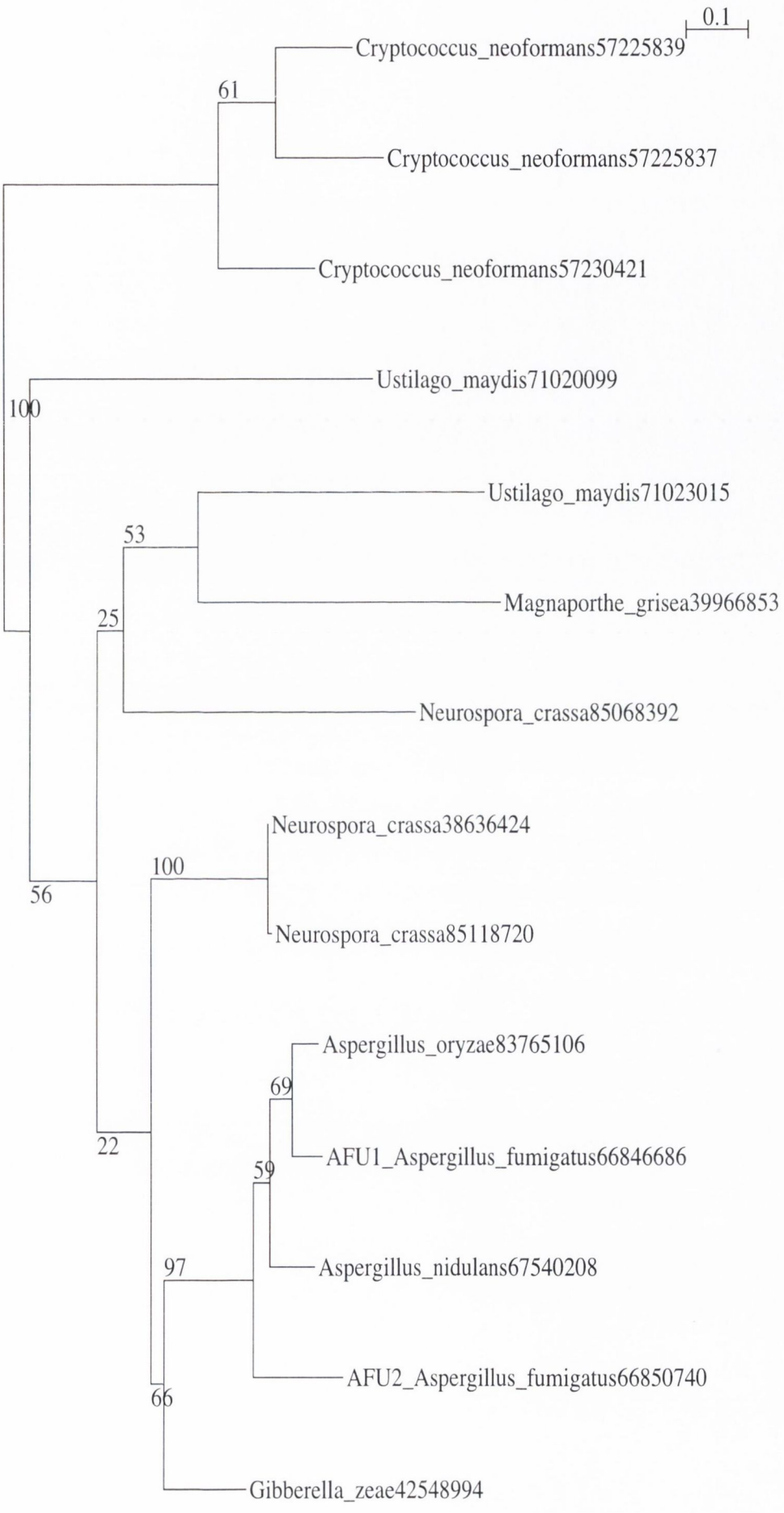
Trees classified as Topology B in *Aspergillus fumigatus* (2).

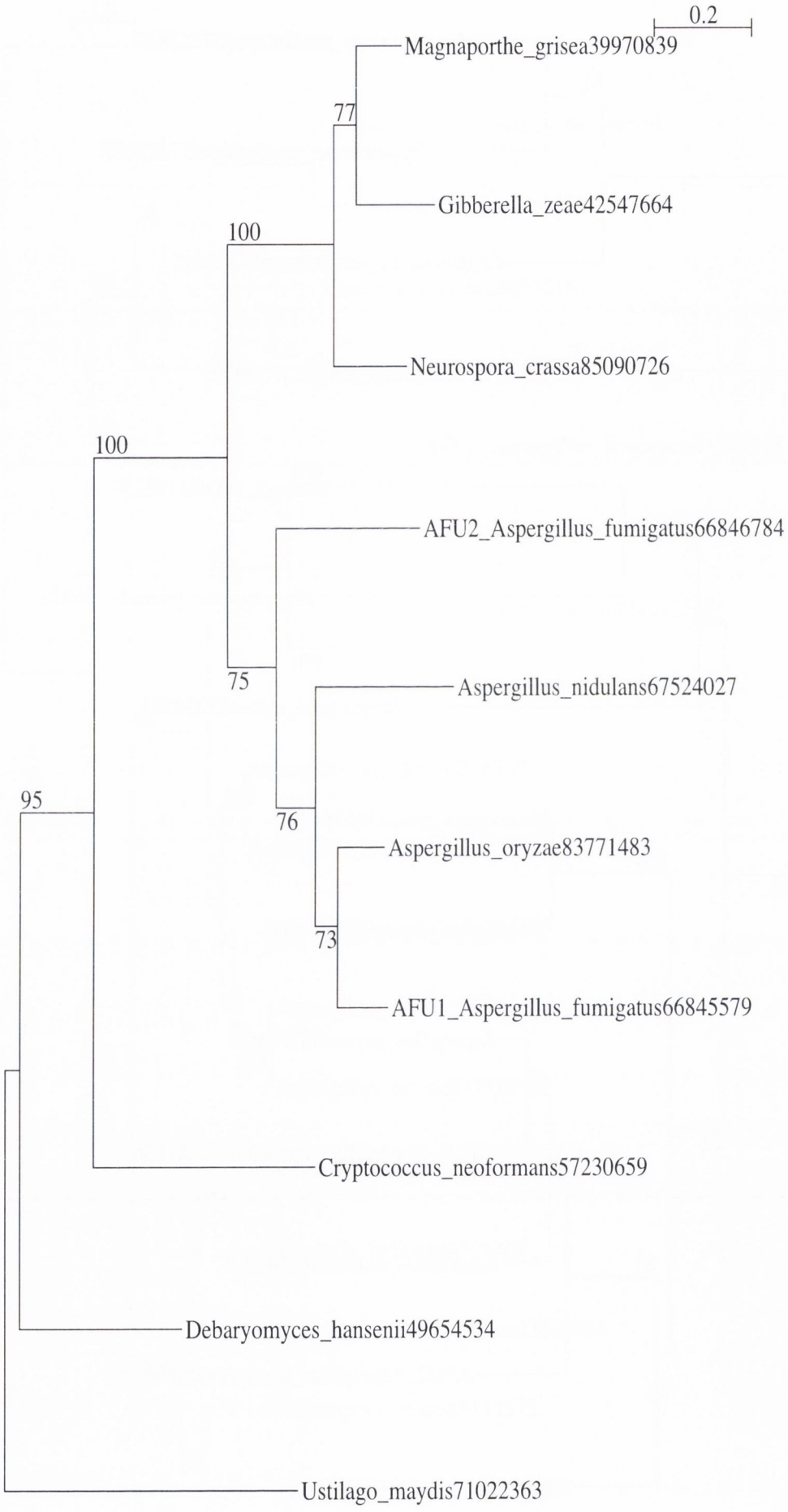

0.2
Ustilago_maydis71006182
Magnaporthe_grisea39943186
84
Neurospora_crassa85074863
AFU2_Aspergillus_fumigatus66846178
Gibberella_zeae42555860
66
Aspergillus_oryzae83773485
24
77
AFU1_Aspergillus_fumigatus66851408
93
Aspergillus_nidulans67537082

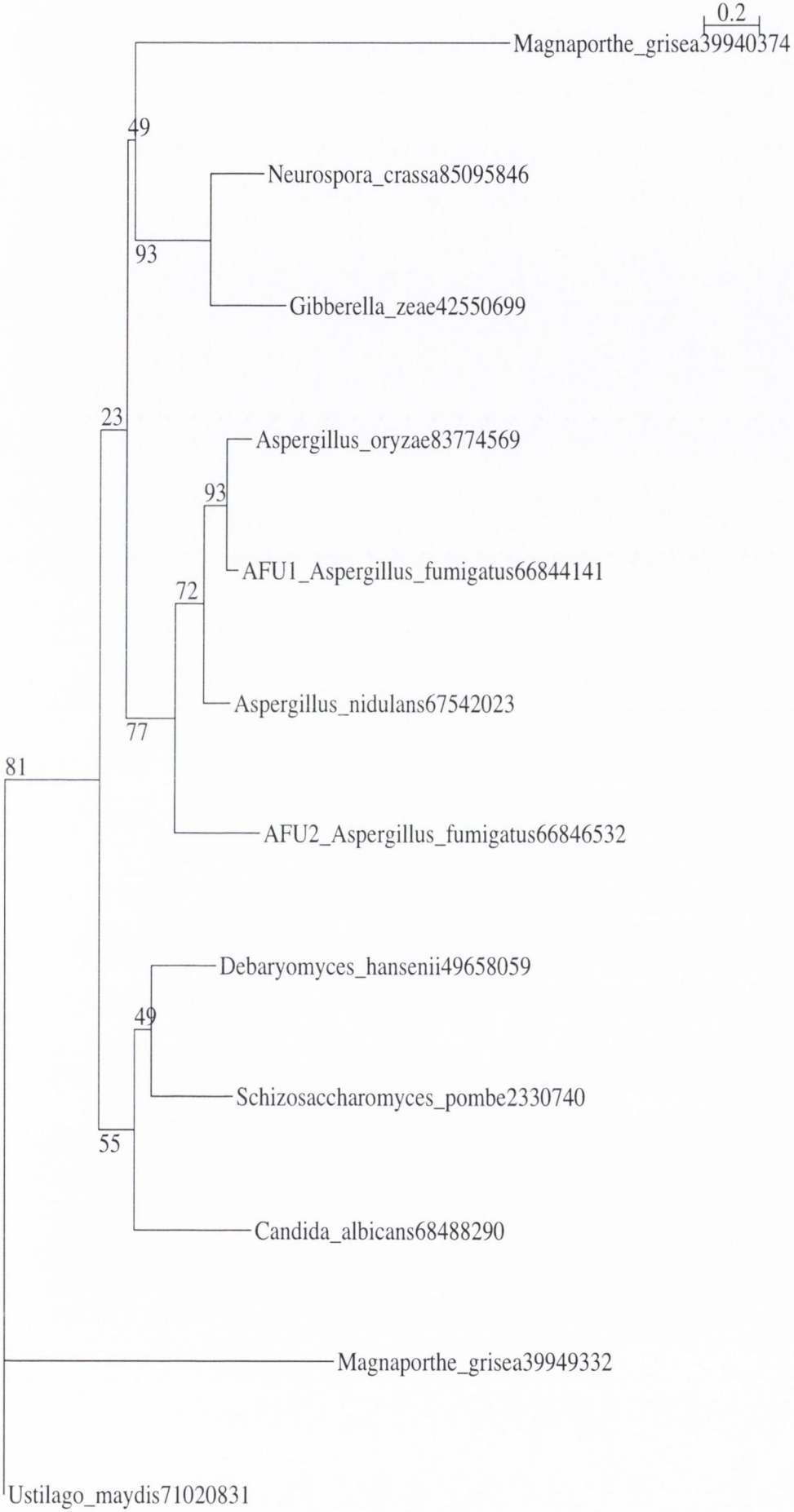
0.2
Ustilago_maydis71018019
Tuber_borchii11066979
AFU2_Aspergillus_fumigatus66846812
92
Aspergillus_fumigatus66845306
Aspergillus_oryzae83767060
69
Aspergillus_niger4140255
68
52
Aspergillus_nidulans67900944
100
AFU1_Aspergillus_fumigatus66847351
65
Hypocrea_jecorina67937782
99
Gibberella_zeae42545889
63
Neurospora_crassa85119610
36
Magnaporthe_grisea39958811

Trees classified as Topology C in *Aspergillus fumigatus* (4).


0.1
Ustilago_maydis71017861
Debaryomyces_hansenii49654539
100
Candida_albicans68484739
96
Yarrowia_lipolytica50552167
63
Schizosaccharomyces_pombe3218399
AFU2_Aspergillus_fumigatus41581201
100
Aspergillus_fumigatus66850123
96
Monascus_pilosus74275567
100
Aspergillus_terreus4959949
96
Aspergillus_nidulans67526719
100
Aspergillus_nidulans29468180
71
Aspergillus_fumigatus66847133
60
Aspergillus_oryzae83772698
69
100
Aspergillus_oryzae83772698
Gibberella_zeae42554194
100
Gibberella_fujikuroi4376229
45
Penicillium_citrinum23574646
100
Neurospora_crassa85113573
60
Magnaporthe_grisea39955070


0.1
Cryptococcus_neoformans57225839
61
Cryptococcus_neoformans57225837
Cryptococcus_neoformans57230421
Ustilago_maydis71020099
100
Ustilago_maydis71023015
53
Magnaporthe_grisea39966853
25
Neurospora_crassa85068392
Neurospora_crassa38636424
100
56
Neurospora_crassa85118720
Aspergillus_oryzae83765106
69
AFU1_Aspergillus_fumigatus66846686
22
59
97
Aspergillus_nidulans67540208
AFU2_Aspergillus_fumigatus66850740
66
Gibberella_zeae42548994


0.2
Magnaporthe_grisea39970839
77
Gibberella_zeae42547664
100
Neurospora_crassa85090726
100
AFU2_Aspergillus_fumigatus66846784
75
Aspergillus_nidulans67524027
95
76
Aspergillus_oryzae83771483
73
AFU1_Aspergillus_fumigatus66845579
Cryptococcus_neoformans57230659
Debaryomyces_hansenii49654534
Ustilago_maydis71022363


0.2
Magnaporthe_grisea39940374
49
Neurospora_crassa85095846
93
Gibberella_zeae42550699
23
Aspergillus_oryzae83774569
93
AFU1_Aspergillus_fumigatus66844141
72
Aspergillus_nidulans67542023
77
81
AFU2_Aspergillus_fumigatus66846532
Debaryomyces_hansenii49658059
49
Schizosaccharomyces_pombe2330740
55
Candida_albicans68488290
Magnaporthe_grisea39949332
Ustilago_maydis71020831

References

Aharonowitz, Y., G. Cohen, and J.F. Martin. 1992. Penicillin and cephalosporin biosynthetic genes: structure, organization, regulation, and evolution. *Annu Rev Microbiol* **46:** 461-495.

Ahmed, A., M.E. Case, and N.H. Giles. 1964. The Nature Of Complementation Among Mutants In The Histidine-3 Region Of Neurospora Crassa. *Brookhaven Symp Biol* **17:** 53-65.

Alberts, A.W., J. Chen, G. Kuron, V. Hunt, J. Huff, C. Hoffman, J. Rothrock, M. Lopez, H. Joshua, E. Harris et al. 1980. Mevinolin: a highly potent competitive inhibitor of hydroxymethylglutaryl-coenzyme A reductase and a cholesterol-lowering agent. *Proc Natl Acad Sci U S A* **77:** 3957-3961.

Amaar, Y.G. and M.M. Moore. 1998. Mapping of the nitrate-assimilation gene cluster (crnA-niiA-niaD) and characterization of the nitrite reductase gene (niiA) in the opportunistic fungal pathogen Aspergillus fumigatus. *Curr Genet* **33:** 206-215.

Ames, B.N., P.E. Hartman, and F. Jacob. 1963. Chromosomal alterations affecting the regulation of histidine biosynthetic enzymes in Salmonella. *J Mol Biol* **7:** 23-42.

An, K.D., H. Nishida, Y. Miura, and A. Yokota. 2003. Molecular evolution of adenylating domain of aminoadipate reductase. *BMC Evol Biol* **3:** 9.

Andersson, J.O., S.W. Sarchfield, and A.J. Roger. 2005. Gene transfers from nanoarchaeota to an ancestor of diplomonads and parabasalids. *Mol Biol Evol* **22:** 85-90.

Andersson, J.O., A.M. Sjogren, L.A. Davis, T.M. Embley, and A.J. Roger. 2003. Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. *Curr Biol* **13:** 94-104.

Andersson, J.O., A.M. Sjogren, D.S. Horner, C.A. Murphy, P.L. Dyal, S.G. Svard, J.M. Logsdon, Jr., M.A. Ragan, R.P. Hirt, and A.J. Roger. 2007. A genomic survey of the fish parasite Spironucleus salmonicida indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution. *BMC Genomics* **8:** 51.

Anonymous. 1999. Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. *Nature* **401:** 921-923.

Anwar, M.M., J.H. Croft, and R.B. Dales. 1993. Analysis of heterokaryon incompatibility between heterokaryon-compatibility (h-c) groups R and GL provides evidence that at least eight het loci control somatic incompatibility in Aspergillus nidulans. *J Gen Microbiol* **139:** 1599-1603.

Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25-29.

Aury, J.M., O. Jaillon, L. Duret, B. Noel, C. Jubin, B.M. Porcel, B. Segurens, V. Daubin, V. Anthouard, N. Aiach et al. 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* **444:** 171-178.

Bai, G.H., A.E. Desjardins, and R.D. Plattner. 2002. Deoxynivalenol-nonproducing fusarium graminearum causes initial infection, but does not cause disease spread in wheat spikes. *Mycopathologia* **153:** 91-98.

Balajee, S.A., S.T. Tay, B.A. Lasker, S.F. Hurst, and A.P. Rooney. 2007. Characterization of a novel gene for strain typing reveals substructuring of Aspergillus fumigatus across North America. *Eukaryot Cell* **6:** 1392-1399.

Banerjee, B., P.A. Greenberger, J.N. Fink, and V.P. Kurup. 1998. Immunological characterization of Asp f 2, a major allergen from Aspergillus fumigatus associated with allergic bronchopulmonary aspergillosis. *Infect Immun* **66:** 5175-5182.

Batada, N.N. and L.D. Hurst. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet* **39:** 945-949.

Batada, N.N., A.O. Urrutia, and L.D. Hurst. 2007. Chromatin remodelling is a major source of coexpression of linked genes in yeast. *Trends Genet* **23:** 480-484.

Beiko, R.G., T.J. Harlow, and M.A. Ragan. 2005. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A* **102:** 14332-14337.

Bentley, R. and J.W. Bennett. 1999. Constructing polyketides: from collie to combinatorial biosynthesis. *Annu Rev Microbiol* **53:** 411-446.

Berriman, M., N. Hall, K. Sheader, F. Bringaud, B. Tiwari, T. Isobe, S. Bowman, C. Corton, L. Clark, G.A. Cross et al. 2002. The architecture of variant surface glycoprotein gene expression sites in Trypanosoma brucei. *Mol Biochem Parasitol* **122:** 131-140.

Bjornerfeldt, S., M.T. Webster, and C. Vila. 2006. Relaxation of selective constraint on dog mitochondrial DNA following domestication. *Genome Res* **16:** 990-994.

Blackwell, M., D.S. Hibbett, J.W. Taylor, and J.W. Spatafora. 2006. Research Coordination Networks: a phylogeny for kingdom Fungi (Deep Hypha). *Mycologia* **98:** 829-837.

Blumenthal, T. 1998. Gene clusters and polycistronic transcription in eukaryotes. *Bioessays* **20:** 480-487.

Blumenthal, T., D. Evans, C.D. Link, A. Guffanti, D. Lawson, J. Thierry-Mieg, D. Thierry-Mieg, W.L. Chiu, K. Duke, M. Kiraly et al. 2002. A global analysis of Caenorhabditis elegans operons. *Nature* **417:** 851-854.

Blumenthal, T. and K.S. Gleason. 2003. Caenorhabditis elegans operons: form and function. *Nat Rev Genet* **4:** 112-120.

Blyth, W., I.W. Grant, E.S. Blackadder, and M. Greenberg. 1977. Fungal antigens as a source of sensitization and respiratory disease in Scottish maltworkers. *Clin Allergy* **7:** 549-562.

Bohnert, H.U., I. Fudal, W. Dioh, D. Tharreau, J.L. Notteghem, and M.H. Lebrun. 2004. A putative polyketide synthase/peptide synthetase from Magnaporthe grisea signals pathogen attack to resistant rice. *Plant Cell* **16:** 2499-2513.

Bojja, R.S., R.L. Cerny, R.H. Proctor, and L. Du. 2004. Determining the biosynthetic sequence in the early steps of the fumonisin pathway by use of three gene-disruption mutants of Fusarium verticillioides. *J Agric Food Chem* **52:** 2855-2860.

Bok, J.W. and N.P. Keller. 2004. LaeA, a regulator of secondary metabolism in Aspergillus spp. *Eukaryot Cell* **3:** 527-535.

Bok, J.W., D. Noordermeer, S.P. Kale, and N.P. Keller. 2006. Secondary metabolic gene cluster silencing in Aspergillus nidulans. *Mol Microbiol* **61:** 1636-1645.

Bortoluzzi, S., L. Rampoldi, B. Simionati, R. Zimbello, A. Barbon, F. d'Alessi, N.

Tiso, A. Pallavicini, S. Toppo, N. Cannata et al. 1998. A comprehensive, high-resolution genomic transcript map of human skeletal muscle. *Genome Res* **8:** 817-825.

Bouhired, S., M. Weber, A. Kempf-Sontag, N.P. Keller, and D. Hoffmeister. 2007. Accurate prediction of the Aspergillus nidulans terrequinone gene cluster boundaries using the transcriptional regulator LaeA. *Fungal Genet Biol.*

Bowyer, P. and D.W. Denning. 2007. Genomic analysis of allergen genes in Aspergillus spp: the relevance of genomics to everyday research. *Med Mycol* **45:** 17-26.

Boylan, M.T., P.M. Mirabito, C.E. Willett, C.R. Zimmerman, and W.E. Timberlake. 1987. Isolation and physical characterization of three essential conidiation genes from Aspergillus nidulans. *Mol Cell Biol* **7:** 3113-3118.

Brown, D.W., T.H. Adams, and N.P. Keller. 1996a. Aspergillus has distinct fatty acid synthases for primary and secondary metabolism. *Proc Natl Acad Sci U S A* **93:** 14873-14877.

Brown, D.W., F. Cheung, R.H. Proctor, R.A. Butchko, L. Zheng, Y. Lee, T. Utterback, S. Smith, T. Feldblyum, A.E. Glenn et al. 2005. Comparative analysis of 87,000 expressed sequence tags from the fumonisin-producing fungus Fusarium verticillioides. *Fungal Genet Biol* **42:** 848-861.

Brown, D.W., R.B. Dyer, S.P. McCormick, D.F. Kendra, and R.D. Plattner. 2004. Functional demarcation of the Fusarium core trichothecene gene cluster. *Fungal Genet Biol* **41:** 454-462.

Brown, D.W., J.H. Yu, H.S. Kelkar, M. Fernandes, T.C. Nesbitt, N.P. Keller, T.H. Adams, and T.J. Leonard. 1996b. Twenty-five coregulated transcripts define a sterigmatocystin gene cluster in Aspergillus nidulans. *Proc Natl Acad Sci U S A* **93:** 1418-1422.

Brown, J.S., A. Aufauvre-Brown, J. Brown, J.M. Jennings, H. Arst, Jr., and D.W. Holden. 2000. Signature-tagged and directed mutagenesis identify PABA synthetase as essential for Aspergillus fumigatus pathogenicity. *Mol Microbiol* **36:** 1371-1380.

Buades, C. and A. Moya. 1996. Phylogenetic analysis of the isopenicillin-N-synthetase horizontal gene transfer. *J Mol Evol* **42:** 537-542.

Burns, R.O., H.E. Umbarger, and S.R. Gross. 1963. The Biosynthesis Of Leucine.

Iii. The Conversion Of Alpha-Hydroxy-Beta-Carboxyisocaproate To Alpha-Ketoisocaproate. *Biochemistry* **2:** 1053-1058.

Byrne, K.P. and K.H. Wolfe. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* **15:** 1456-1461.

Cai, J.J., P.C. Woo, S.K. Lau, D.K. Smith, and K.Y. Yuen. 2006. Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota. *J Mol Evol* **63:** 1-11.

Calvo, A.M., H.W. Gardner, and N.P. Keller. 2001. Genetic connection between fatty acid metabolism and sporulation in Aspergillus nidulans. *J Biol Chem* **276:** 25766-25774.

Calvo, A.M., R.A. Wilson, J.W. Bok, and N.P. Keller. 2002. Relationship between secondary metabolism and fungal development. *Microbiol Mol Biol Rev* **66:** 447-459, table of contents.

Caron, H., B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M.C. Hermus, R. van Asperen, K. Boon, P.A. Voute et al. 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291:** 1289-1292.

Cary, J.W., P.-K. Chang, and D. Bhatnagar. 2001. *Clustered metabolic pathway genes in filamentous fungi.* In Applied Mycology and Biotechnology. Edited by: Khachatourians, G.G. Arora, D.K., Amsterdam: Elsevier.

Cary, J.W., J.M. Dyer, K.C. Ehrlich, M.S. Wright, S.H. Liang, and J.E. Linz. 2002. Molecular and functional characterization of a second copy of the aflatoxin regulatory gene, aflR-2, from Aspergillus parasiticus. *Biochim Biophys Acta* **1576:** 316-323.

Cary, J.W. and K.C. Ehrlich. 2006. Aflatoxigenicity in Aspergillus: molecular genetics, phylogenetic relationships and evolutionary implications. *Mycopathologia* **162:** 167-177.

Casadevall, A., J.N. Steenbergen, and J.D. Nosanchuk. 2003. 'Ready made' virulence and 'dual use' virulence factors in pathogenic environmental fungi--the Cryptococcus neoformans paradigm. *Curr Opin Microbiol* **6:** 332-337.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17:** 540-552.

Champe, S.P. and A.A. el-Zayat. 1989. Isolation of a sexual sporulation hormone from Aspergillus nidulans. *J Bacteriol* **171:** 3982-3988.

Champe, S.P., P. Rao, and A. Chang. 1987. An endogenous inducer of sexual development in Aspergillus nidulans. *J Gen Microbiol* **133:** 1383-1387.

Chang, P.K., K.C. Ehrlich, J.E. Linz, D. Bhatnagar, T.E. Cleveland, and J.W. Bennett. 1996. Characterization of the Aspergillus parasiticus niaD and niiA gene cluster. *Curr Genet* **30:** 68-75.

Chang, P.K. and J. Yu. 2002. Characterization of a partial duplication of the aflatoxin gene cluster in Aspergillus parasiticus ATCC 56775. *Appl Microbiol Biotechnol* **58:** 632-636.

Changa, P.-K., K.C. Ehrlich, and S.-S.T. Hua. 2006. Cladal relatedness among Aspergillus oryzae isolates and Aspergillus flavus S and L morphotype isolates. *International Journal of Food Microbiology* **108:** 172-177.

Chen, F.C. and W.H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* **68:** 444-456.

Chim, C.S., P.L. Ho, and K.Y. Yuen. 1998. Simultaneous Aspergillus fischeri and Herpes simplex pneumonia in a patient with multiple myeloma. *Scand J Infect Dis* **30:** 190-191.

Choi, I.G. and S.H. Kim. 2007. Global extent of horizontal gene transfer. *Proc Natl Acad Sci U S A* **104:** 4489-4494.

Clemons, K.V., T.K. Miller, C.P. Selitrennikoff, and D.A. Stevens. 2002. fos-1, a putative histidine kinase as a virulence factor for systemic aspergillosis. *Med Mycol* **40:** 259-262.

Cohen, B.A., R.D. Mitra, J.D. Hughes, and G.M. Church. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* **26:** 183-186.

Collemare, J., A. Billard, H.U. Bohnert, and M.H. Lebrun. 2008. Biosynthesis of secondary metabolites in the rice blast fungus Magnaporthe grisea: the role of hybrid PKS-NRPS in pathogenicity. *Mycol Res* **112:** 207-215.

Conant, G.C. and A. Wagner. 2003. Asymmetric sequence divergence of duplicate genes. *Genome Res* **13:** 2052-2058.

Coyle, C.M. and D.G. Panaccione. 2005. An ergot alkaloid biosynthesis gene and clustered hypothetical genes from Aspergillus fumigatus. *Appl Environ*

Microbiol **71:** 3112-3118.

Cuomo, C.A., U. Guldener, J.R. Xu, F. Trail, B.G. Turgeon, A. Di Pietro, J.D. Walton, L.J. Ma, S.E. Baker, M. Rep et al. 2007. The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. *Science* **317:** 1400-1402.

Czaran, T.L., R.F. Hoekstra, and L. Pagie. 2002. Chemical warfare between microbes promotes biodiversity. *Proc Natl Acad Sci U S A* **99:** 786-790.

d'Enfert, C. 1996. Selection of multiple disruption events in Aspergillus fumigatus using the orotidine-5'-decarboxylase gene, pyrG, as a unique transformation marker. *Curr Genet* **30:** 76-82.

Darwin, C.R. 1859. *On the Origin of Species by Means of Natural Selection.* John Murray, London.

Davies, G. and B. Henrissat. 1995. Structures and mechanisms of glycosyl hydrolases. *Structure* **3:** 853-859.

Dean, R.A., N.J. Talbot, D.J. Ebbole, M.L. Farman, T.K. Mitchell, M.J. Orbach, M. Thon, R. Kulkarni, J.R. Xu, H. Pan et al. 2005. The genome sequence of the rice blast fungus Magnaporthe grisea. *Nature* **434:** 980-986.

Dehal, P. and J.L. Boore. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3:** e314.

Delcher, A.L., S. Kasif, R.D. Fleischmann, J. Peterson, O. White, and S.L. Salzberg. 1999. Alignment of whole genomes. *Nucleic Acids Res* **27:** 2369-2376.

Demerec, M. and H. Ozeki. 1959. Tests for Allelism among Auxotrophs of Salmonella Typhimurium. *Genetics* **44:** 269-278.

Denning, D.W. 1998. Invasive aspergillosis. *Clin Infect Dis* **26:** 781-803; quiz 804-785.

Diez, B., S. Gutierrez, J.L. Barredo, P. van Solingen, L.H. van der Voort, and J.F. Martin. 1990. The cluster of penicillin biosynthetic genes. Identification and characterization of the pcbAB gene encoding the alpha-aminoadipyl-cysteinyl-valine synthetase and linkage to the pcbC and penDE genes. *J Biol Chem* **265:** 16358-16365.

Doebley, J.F., B.S. Gaut, and B.D. Smith. 2006. The molecular genetics of crop domestication. *Cell* **127:** 1309-1321.

Doolittle, W.F. 1999. Phylogenetic classification and the universal tree. *Science*

284: 2124-2129.

Du, L., Y.Q. Cheng, G. Ingenhorst, G.L. Tang, Y. Huang, and B. Shen. 2003. Hybrid peptide-polyketide natural products: biosynthesis and prospects towards engineering novel molecules. *Genet Eng (N Y)* **25:** 227-267.

Du, L., C. Sanchez, and B. Shen. 2001. Hybrid peptide-polyketide natural products: biosynthesis and prospects toward engineering novel molecules. *Metab Eng* **3:** 78-95.

Du, L. and B. Shen. 2001. Biosynthesis of hybrid peptide-polyketide natural products. *Curr Opin Drug Discov Devel* **4:** 215-228.

Dudareva, N., S. Andersson, I. Orlova, N. Gatto, M. Reichelt, D. Rhodes, W. Boland, and J. Gershenzon. 2005. The nonmevalonate pathway supports both monoterpene and sesquiterpene formation in snapdragon flowers. *Proc Natl Acad Sci U S A* **102:** 933-938.

Duo-Chuan, L. 2006. Review of fungal chitinases. *Mycopathologia* **161:** 345-360.

Duroux, L., F.M. Delmotte, J.M. Lancelin, G. Keravis, and C. Jay-Allemand. 1998. Insight into naphthoquinone metabolism: beta-glucosidase-catalysed hydrolysis of hydrojuglone beta-D-glucopyranoside. *Biochem J* **333 (Pt 2):** 275-283.

Dyer, P.S. and M. Paoletti. 2005. Reproduction in Aspergillus fumigatus: sexuality in a supposedly asexual species? *Med Mycol* **43 Suppl 1:** S7-14.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792-1797.

Ehrlich, K.C. and P.J. Cotty. 2004. An isolate of Aspergillus flavus used to reduce aflatoxin contamination in cottonseed has a defective polyketide synthase gene. *Appl Microbiol Biotechnol* **65:** 473-478.

Eisendle, M., H. Oberegger, I. Zadra, and H. Haas. 2003. The siderophore system is essential for viability of Aspergillus nidulans: functional analysis of two genes encoding l-ornithine N 5-monooxygenase (sidA) and a non-ribosomal peptide synthetase (sidC). *Mol Microbiol* **49:** 359-375.

Fares, M.A., K.P. Byrne, and K.H. Wolfe. 2006. Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of Saccharomyces species. *Mol Biol Evol* **23:** 245-253.

Faure, D., B. Henrissat, D. Ptacek, M.A. Bekri, and J. Vanderleyden. 2001. The celA gene, encoding a glycosyl hydrolase family 3 beta-glucosidase in

Azospirillum irakense, is required for optimal growth on cellobiosides. *Appl Environ Microbiol* **67:** 2380-2383.

Fawcett, H.S. 1925. Maintained growth rates in fungus cultures of long duration. *Ann. Appl. Biol.* **12:** 191-198.

Fedorova, N.D., J.H. Badger, G.D. Robson, J.R. Wortman, and W.C. Nierman. 2005. Comparative analysis of programmed cell death pathways in filamentous fungi. *BMC Genomics* **6:** 177.

Felenbok, B. 1991. The ethanol utilization regulon of Aspergillus nidulans: the alcA-alcR system as a tool for the expression of recombinant proteins. *J Biotechnol* **17:** 11-17.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. *Syst Zool* **27:** 401-410.

Field, B. and A.E. Osbourn. 2008. Metabolic Diversification--Independent Assembly of Operon-Like Gene Clusters in Plants. *Science* **320:** 543-547.

Fillinger, S. and B. Felenbok. 1996. A newly identified gene cluster in Aspergillus nidulans comprises five novel genes localized in the alc region that are controlled both by the specific transactivator AlcR and the general carbon-catabolite repressor CreA. *Mol Microbiol* **20:** 475-488.

Finking, R. and M.A. Marahiel. 2004. Biosynthesis of nonribosomal peptides1. *Annu Rev Microbiol* **58:** 453-488.

Fischer, G., E.P. Rocha, F. Brunet, M. Vergassola, and B. Dujon. 2006. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS Genet* **2:** e32.

Fitzpatrick, D.A., M.E. Logue, J.E. Stajich, and G. Butler. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol Biol* **6:** 99.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269:** 496-512.

Fowler, T. 1993. Deletion of the Trichoderma reesei Beta-glucosidase gene, bgl1. In *ACS Symposium Series 533*, pp. 56-65. Edited by: Esen A. Washington DC: American Chemical Society.

Friesen, T.L., E.H. Stukenbrock, Z. Liu, S. Meinhardt, H. Ling, J.D. Faris, J.B.

Rasmussen, P.S. Solomon, B.A. McDonald, and R.P. Oliver. 2006. Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat Genet* **38:** 953-956.

Frigaard, N.U., A. Martinez, T.J. Mincer, and E.F. DeLong. 2006. Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439:** 847-850.

Frisvad, J.C. and R.A. Samson. 2004. Emericella venezuelensis, a new species with stellate ascospores producing sterigmatocystin and aflatoxin B1. *Syst Appl Microbiol* **27:** 672-680.

Fukuoka, Y., H. Inaoka, and I.S. Kohane. 2004. Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. *BMC Genomics* **5:** 4.

Galagan, J.E., S.E. Calvo, K.A. Borkovich, E.U. Selker, N.D. Read, D. Jaffe, W. FitzHugh, L.J. Ma, S. Smirnov, S. Purcell et al. 2003. The genome sequence of the filamentous fungus Neurospora crassa. *Nature* **422:** 859-868.

Galagan, J.E., S.E. Calvo, C. Cuomo, L.J. Ma, J.R. Wortman, S. Batzoglou, S.I. Lee, M. Basturkmen, C.C. Spevak, J. Clutterbuck et al. 2005. Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature* **438:** 1105-1115.

Garcia-Vallve, S., A. Romeu, and J. Palau. 2000. Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. *Mol Biol Evol* **17:** 352-361.

Gardiner, D.M., A.J. Cozijnsen, L.M. Wilson, M.S. Pedras, and B.J. Howlett. 2004. The sirodesmin biosynthetic gene cluster of the plant pathogenic fungus Leptosphaeria maculans. *Mol Microbiol* **53:** 1307-1318.

Gardiner, D.M. and B.J. Howlett. 2005. Bioinformatic and expression analysis of the putative gliotoxin biosynthetic gene cluster of Aspergillus fumigatus. *FEMS Microbiol Lett* **248:** 241-248.

Gardiner, D.M., R.S. Jarvis, and B.J. Howlett. 2005a. The ABC transporter gene in the sirodesmin biosynthetic gene cluster of Leptosphaeria maculans is not essential for sirodesmin production but facilitates self-protection. *Fungal Genet Biol* **42:** 257-263.

Gardiner, D.M., P. Waring, and B.J. Howlett. 2005b. The epipolythiodioxopiperazine (ETP) class of fungal toxins: distribution,

mode of action, functions and biosynthesis. *Microbiology* **151:** 1021-1032.

Geiser, D.M., J.I. Pitt, and J.W. Taylor. 1998. Cryptic speciation and recombination in the aflatoxin-producing fungus Aspergillus flavus. *Proc Natl Acad Sci U S A* **95:** 388-393.

Gierl, A. and M. Frey. 2001. Evolution of benzoxazinone biosynthesis and indole production in maize. *Planta* **213:** 493-498.

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

Glass, N.L. and K. Dementhon. 2006. Non-self recognition and programmed cell death in filamentous fungi. *Curr Opin Microbiol* **9:** 553-558.

Goffeau, A. 2005. Genomics: multiple moulds. *Nature* **438:** 1092-1093.

Gottlieb, D. 1971. Limited growth of fungi. *Mycologia* **63:** 619-629.

Graham, A., N. Papalopulu, and R. Krumlauf. 1989. The murine and Drosophila homeobox gene complexes have common features of organization and expression. *Cell* **57:** 367-378.

Gross, S.R., R.O. Burns, and H.E. Umbarger. 1963. The Biosynthesis Of Leucine. Ii. The Enzymic Isomerization Of Beta-Carboxy-Beta-Hydroxyisocaproate And Alpha-Hydroxy-Beta-Carboxyisocaproate. *Biochemistry* **2:** 1046-1052.

Gu, Z., L. David, D. Petrov, T. Jones, R.W. Davis, and L.M. Steinmetz. 2005. Elevated evolutionary rates in the laboratory strain of Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **102:** 1092-1097.

Guindon, S. and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52:** 696-704.

Guljamow, A., H. Jenke-Kodama, H. Saumweber, P. Quillardet, L. Frangeul, A.M. Castets, C. Bouchier, N. Tandeau de Marsac, and E. Dittmann. 2007. Horizontal gene transfer of two cytoskeletal elements from a eukaryote to a cyanobacterium. *Curr Biol* **17:** R757-759.

Gwynne, D.I., B.L. Miller, K.Y. Miller, and W.E. Timberlake. 1984. Structure and regulated expression of the SpoC1 gene cluster from Aspergillus nidulans. *J Mol Biol* **180:** 91-109.

Haarmann, T., I. Ortel, P. Tudzynski, and U. Keller. 2006. Identification of the cytochrome P450 monooxygenase that bridges the clavine and ergoline

alkaloid pathways. *Chembiochem* **7:** 645-652.

Haas, B.J., A.L. Delcher, S.M. Mount, J.R. Wortman, R.K. Smith, Jr., L.I. Hannick, R. Maiti, C.M. Ronning, D.B. Rusch, C.D. Town et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31:** 5654-5666.

Haas, B.J., S.L. Salzberg, W. Zhu, M. Pertea, J.E. Allen, J. Orvis, O. White, C.R. Buell, and J.R. Wortman. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9:** R7.

Hall, C., S. Brachat, and F.S. Dietrich. 2005. Contribution of horizontal gene transfer to the evolution of Saccharomyces cerevisiae. *Eukaryot Cell* **4:** 1102-1115.

Hall, C. and F.S. Dietrich. 2007. The reacquisition of biotin prototrophy in Saccharomyces cerevisiae involved horizontal gene transfer, gene duplication and gene clustering. *Genetics* **177:** 2293-2307.

Hamada, T., Y. Fukushima, and H. Motai. 1992. *Continuous production of soy sauce in a bioreactor. Applications of Biotechnology to Traditional Fermented Foods* Report of an Ad Hoc panel of the Board on Science and Technology for International Development Office of International Affairs, Washington DC: National Academy Press. pp. 114-118.

Han, K.H., J.A. Seo, and J.H. Yu. 2004. Regulators of G-protein signalling in Aspergillus nidulans: RgsA downregulates stress response and stimulates asexual sporulation through attenuation of GanB (Galpha) signalling. *Mol Microbiol* **53:** 529-540.

Hane, J.K., R.G. Lowe, P.S. Solomon, K.C. Tan, C.L. Schoch, J.W. Spatafora, P.W. Crous, C. Kodira, B.W. Birren, J.E. Galagan et al. 2007. Dothideomycete plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen Stagonospora nodorum. *Plant Cell* **19:** 3347-3368.

Haralampidis, K., G. Bryan, X. Qi, K. Papadopoulou, S. Bakht, R. Melton, and A. Osbourn. 2001. A new class of oxidosqualene cyclases directs synthesis of antimicrobial phytoprotectants in monocots. *Proc Natl Acad Sci U S A* **98:** 13431-13436.

Henrissat, B. and G. Davies. 1997. Structural and sequence-based classification of

glycoside hydrolases. *Curr Opin Struct Biol* **7:** 637-644.

Henson, J.M., M.J. Butler, and A.W. Day. 1999. The Dark Side of the Mycelium: Melanins of Phytopathogenic Fungi. *Annu Rev Phytopathol* **37:** 447-471.

Hibbett, D.S., M. Binder, J.F. Bischoff, M. Blackwell, P.F. Cannon, O.E. Eriksson, S. Huhndorf, T. James, P.M. Kirk, R. Lucking et al. 2007. A higher-level phylogenetic classification of the Fungi. *Mycol Res* **111:** 509-547.

Hoffmeister, D. and N.P. Keller. 2007. Natural products of filamentous fungi: enzymes, genes, and their regulation. *Nat Prod Rep* **24:** 393-416.

Hohn, T.M. and P.D. Beremand. 1989. Isolation and nucleotide sequence of a sesquiterpene cyclase gene from the trichothecene-producing fungus Fusarium sporotrichioides. *Gene* **79:** 131-138.

Hohn, T.M. and A.E. Desjardins. 1992. Isolation and gene disruption of the Tox5 gene encoding trichodiene synthase in Gibberella pulicaris. *Mol Plant Microbe Interact* **5:** 249-256.

Hopwood, D.A. 1997. Genetic Contributions to Understanding Polyketide Synthases. *Chem Rev* **97:** 2465-2498.

Hopwood, D.A. and D.H. Sherman. 1990. Molecular genetics of polyketides and its comparison to fatty acid biosynthesis. *Annu Rev Genet* **24:** 37-66.

Horowitz, N.H. 1965. The Evolution of Biochemical Sytheses- Retrospect and Prospect. In *Evolving genes and proteins* eds V. Bryson and B.J. Vogel). Academic Press, Institute of Microbiology of Rutgers.

Horowitzn, H. 1945. On the evolution of biochemical synthesis. *Proc. Natl. Acad. Sci.*

31: 153-157.

Hosack, D.A., G. Dennis, Jr., B.T. Sherman, H.C. Lane, and R.A. Lempicki. 2003. Identifying biological themes within lists of genes with EASE. *Genome Biol* **4:** R70.

Hrmova, M., M. Banik, A.J. Harvey, T.P. Garrett, J.N. Varghese, P.B. Hoj, and G.B. Fincher. 1997. Polysaccharide hydrolases in germinated barley and their role in the depolymerization of plant and fungal cell walls. *Int J Biol Macromol* **21:** 67-72.

Hull, E.P., P.M. Green, H.N. Arst, Jr., and C. Scazzocchio. 1989. Cloning and physical characterization of the L-proline catabolism gene cluster of

Aspergillus nidulans. *Mol Microbiol* **3:** 553-559.

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

Hurst, L.D., E.J. Williams, and C. Pal. 2002. Natural selection promotes the conservation of linkage of co-expressed genes. *Trends Genet* **18:** 604-606.

Hutchinson, C.R., J. Kennedy, C. Park, S. Kendrew, K. Auclair, and J. Vederas. 2000. Aspects of the biosynthesis of non-aromatic fungal polyketides by iterative polyketide synthases. *Antonie Van Leeuwenhoek* **78:** 287-295.

Jaillon, O., J.M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449:** 463-467.

James, T.Y., F. Kauff, C.L. Schoch, P.B. Matheny, V. Hofstetter, C.J. Cox, G. Celio, C. Gueidan, E. Fraker, J. Miadlikowska et al. 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443:** 818-822.

Jansen, C., D. von Wettstein, W. Schafer, K.H. Kogel, A. Felk, and F.J. Maier. 2005. Infection patterns in barley and wheat spikes inoculated with wild-type and trichodiene synthase gene disrupted Fusarium graminearum. *Proc Natl Acad Sci U S A* **102:** 16892-16897.

Jeffries, T.W., I.V. Grigoriev, J. Grimwood, J.M. Laplaza, A. Aerts, A. Salamov, J. Schmutz, E. Lindquist, P. Dehal, H. Shapiro et al. 2007. Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast Pichia stipitis. *Nat Biotechnol* **25:** 319-326.

Johnstone, I.L., P.C. McCabe, P. Greaves, S.J. Gurr, G.E. Cole, M.A. Brow, S.E. Unkles, A.J. Clutterbuck, J.R. Kinghorn, and M.A. Innis. 1990. Isolation and characterisation of the crnA-niiA-niaD gene cluster for nitrate assimilation in Aspergillus nidulans. *Gene* **90:** 181-192.

Jungwirth, C., S.R. Gross, P. Margolin, and H.E. Umbarger. 1963. The biosynthesis of leucine. I. The accumulation of beta-carboxy-beta-hydroxyisocaproate by leucine auxotrophs of Salmonella typhimurium and Neurospora crassa. *Biochemistry* **2:** 1-6.

Kale, S.P., J.W. Cary, D. Bhatnagar, and J.W. Bennett. 1996. Characterization of experimentally induced, nonaflatoxigenic variant strains of Aspergillus

parasiticus. *Appl Environ Microbiol* **62:** 3399-3404.

Kamper, J., R. Kahmann, M. Bolker, L.J. Ma, T. Brefort, B.J. Saville, F. Banuett, J.W. Kronstad, S.E. Gold, O. Muller et al. 2006. Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis. *Nature* **444:** 97-101.

Kawamura, C., T. Tsujimoto, and T. Tsuge. 1999. Targeted disruption of a melanin biosynthesis gene affects conidial development and UV tolerance in the Japanese pear pathotype of Alternaria alternata. *Mol Plant Microbe Interact* **12:** 59-63.

Keller, N., J. Bok, D. Chung, R.M. Perrin, and E. Keats Shwab. 2006. LaeA, a global regulator of Aspergillus toxins. *Med Mycol* **44 Suppl:** 83-85.

Keller, N.P. and T.M. Hohn. 1997. Metabolic Pathway Gene Clusters in Filamentous Fungi. *Fungal Genet Biol* **21:** 17-29.

Keller, N.P., G. Turner, and J.W. Bennett. 2005. Fungal secondary metabolism - from biochemistry to genomics. *Nat Rev Microbiol* **3:** 937-947.

Kellerman, T.S., J.G. Pienaar, G.C. van der Westhuizen, G.C. Anderson, and T.W. Naude. 1976. A highly fatal tremorgenic mycotoxicosis of cattle caused by Aspergillus clavatus. *Onderstepoort J Vet Res* **43:** 147-154.

Kellis, M., B.W. Birren, and E.S. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428:** 617-624.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E.S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241-254.

Kennedy, J., K. Auclair, S.G. Kendrew, C. Park, J.C. Vederas, and C.R. Hutchinson. 1999. Modulation of polyketide synthase activity by accessory proteins during lovastatin biosynthesis. *Science* **284:** 1368-1372.

Keszenman-Pereyra, D., S. Lawrence, M.E. Twfieg, J. Price, and G. Turner. 2003. The npgA/ cfwA gene encodes a putative 4'-phosphopantetheinyl transferase which is essential for penicillin biosynthesis in Aspergillus nidulans. *Curr Genet* **43:** 186-190.

Khaldi, N., J. Collemare, M.H. Lebrun, and K.H. Wolfe. 2008. Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. *Genome Biol* **9:** R18.

Kimura, M. and T. Ohta. 1974. On some principles governing molecular evolution. *Proc Natl Acad Sci U S A* **71:** 2848-2852.

Kimura, M., T. Tokai, K. O'Donnell, T.J. Ward, M. Fujimura, H. Hamamoto, T. Shibata, and I. Yamaguchi. 2003. The trichothecene biosynthesis gene cluster of Fusarium graminearum F15 contains a limited number of essential pathway genes and expressed non-essential genes. *FEBS Lett* **539:** 105-110.

Kobayashi, T., K. Abe, K. Asai, K. Gomi, P.R. Juvvadi, M. Kato, K. Kitamoto, M. Takeuchi, and M. Machida. 2007. Genomics of Aspergillus oryzae. *Biosci Biotechnol Biochem* **71:** 646-670.

Kodama, M., M.S. Rose, G. Yang, S.H. Yun, O.C. Yoder, and B.G. Turgeon. 1999. The translocation-associated tox1 locus of Cochliobolus heterostrophus is two genetic elements on two different chromosomes. *Genetics* **151:** 585-596.

Kondo, N., N. Nikoh, N. Ijichi, M. Shimada, and T. Fukatsu. 2002. Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect. *Proc Natl Acad Sci U S A* **99:** 14280-14285.

Kroken, S., N.L. Glass, J.W. Taylor, O.C. Yoder, and B.G. Turgeon. 2003. Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proc Natl Acad Sci U S A* **100:** 15670-15675.

Kruglyak, S. and H. Tang. 2000. Regulation of adjacent yeast genes. *Trends Genet* **16:** 109-111.

Kuramae, E.E., V. Robert, B. Snel, M. Weiss, and T. Boekhout. 2006. Phylogenomics reveal a robust fungal tree of life. *FEMS Yeast Res* **6:** 1213-1220.

Kurland, C.G., B. Canback, and O.G. Berg. 2003. Horizontal gene transfer: a critical view. *Proc Natl Acad Sci U S A* **100:** 9658-9662.

Langfelder, K., B. Jahn, H. Gehringer, A. Schmidt, G. Wanner, and A.A. Brakhage. 1998. Identification of a polyketide synthase gene (pksP) of Aspergillus fumigatus involved in conidial pigment biosynthesis and virulence. *Med Microbiol Immunol* **187:** 79-89.

Langkjaer, R.B., P.F. Cliften, M. Johnston, and J. Piskur. 2003. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. *Nature* **421:** 848-852.

Latge, J.P. 1999. Aspergillus fumigatus and aspergillosis. *Clin Microbiol Rev* **12:** 310-350.

Lawrence, J. 1999. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr Opin Genet Dev* **9:** 642-648.

Lawrence, J.G. and J.R. Roth. 1996. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143:** 1843-1860.

Lee, D.G., J.M. Urbach, G. Wu, N.T. Liberati, R.L. Feinbaum, S. Miyata, L.T. Diggins, J. He, M. Saucier, E. Deziel et al. 2006. Genomic analysis reveals that Pseudomonas aeruginosa virulence is combinatorial. *Genome Biol* **7:** R90.

Lee, D.Y., C. Teyssier, B.D. Strahl, and M.R. Stallcup. 2005. Role of protein methylation in regulation of transcription. *Endocr Rev* **26:** 147-170.

Lee, J.M. and E.L. Sonnhammer. 2003. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res* **13:** 875-882.

Leipe, D.D., E.V. Koonin, and L. Aravind. 2004. STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer. *J Mol Biol* **343:** 1-28.

Lercher, M.J., T. Blumenthal, and L.D. Hurst. 2003. Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. *Genome Res* **13:** 238-243.

Lercher, M.J., A.O. Urrutia, and L.D. Hurst. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* **31:** 180-183.

Lester, G. 1963. Regulation of early reactions in the biosynthesis of tryptophan in Neurospora crassa. *J Bacteriol* **85:** 468-475.

Lindblad-Toh, K. C.M. Wade T.S. Mikkelsen E.K. Karlsson D.B. Jaffe M. Kamal M. Clamp J.L. Chang E.J. Kulbokas, 3rd M.C. Zody et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438:** 803-819.

Liras, P. and J.F. Martin. 2006. Gene clusters for beta-lactam antibiotics and control of their expression: why have clusters evolved, and from where did they originate? *Int Microbiol* **9:** 9-19.

Liu, Y.J. and B.D. Hall. 2004. Body plan evolution of ascomycetes, as inferred

from an RNA polymerase II phylogeny. *Proc Natl Acad Sci U S A* **101:** 4507-4512.

Lonial, S., L. Williams, G. Carrum, M. Ostrowski, and P. McCarthy, Jr. 1997. Neosartorya fischeri: an invasive fungal pathogen in an allogeneic bone marrow transplant patient. *Bone Marrow Transplant* **19:** 753-755.

Lutz, M.P., G. Feichtinger, G. Defago, and B. Duffy. 2003. Mycotoxigenic Fusarium and deoxynivalenol production repress chitinase gene expression in the biocontrol agent Trichoderma atroviride P1. *Appl Environ Microbiol* **69:** 3077-3084.

Lynch, M. and J.S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151-1155.

Machida, M., K. Asai, M. Sano, T. Tanaka, T. Kumagai, G. Terai, K. Kusumoto, T. Arima, O. Akita, Y. Kashiwagi et al. 2005. Genome sequencing and analysis of Aspergillus oryzae. *Nature* **438:** 1157-1161.

Maiya, S., A. Grundmann, S.M. Li, and G. Turner. 2006. The fumitremorgin gene cluster of Aspergillus fumigatus: identification of a gene encoding brevianamide F synthetase. *Chembiochem* **7:** 1062-1069.

Maiya, S., A. Grundmann, X. Li, S.M. Li, and G. Turner. 2007. Identification of a hybrid PKS/NRPS required for pseurotin A biosynthesis in the human pathogen Aspergillus fumigatus. *Chembiochem* **8:** 1736-1743.

Malz, S., M.N. Grell, C. Thrane, F.J. Maier, P. Rosager, A. Felk, K.S. Albertsen, S. Salomon, L. Bohn, W. Schafer et al. 2005. Identification of a gene cluster responsible for the biosynthesis of aurofusarin in the Fusarium graminearum species complex. *Fungal Genet Biol* **42:** 420-433.

Marahiel, M.A., T. Stachelhaus, and H.D. Mootz. 1997. Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. *Chem Rev* **97:** 2651-2674.

Margolin, P. 1963. Genetic fine structure of the leucine operon in Salmonella. *Genetics* **48:** 441-457.

Marquez-Fernandez, O., A. Trigos, J.L. Ramos-Balderas, G. Viniegra-Gonzalez, H.B. Deising, and J. Aguirre. 2007. Phosphopantetheinyl transferase CfwA/NpgA is required for Aspergillus nidulans secondary metabolism and asexual development. *Eukaryot Cell* **6:** 710-720.

McDonald, T., D. Noordermeer, Y.-Q. Zhang, T. Hammond, and N.P. Keller.

2005. *The ST cluster revisited: lessons from a genetic model*. Marcel Dekker Inc, New York.

Mefford, H.C. and B.J. Trask. 2002. The complex structure and dynamic evolution of human subtelomeres. *Nat Rev Genet* **3:** 91-102.

Megy, K., S. Audic, and J.M. Claverie. 2003. Positional clustering of differentially expressed genes on human chromosomes 20, 21 and 22. *Genome Biol* **4:** P1.

Merrick, C.J. and M.T. Duraisingh. 2006. Heterochromatin-mediated control of virulence gene expression. *Mol Microbiol* **62:** 612-620.

Miller, D.A. and C.T. Walsh. 2001. Yersiniabactin synthetase: probing the recognition of carrier protein domains by the catalytic heterocyclization domains, Cy1 and Cy2, in the chain-initiating HWMP2 subunit. *Biochemistry* **40:** 5313-5321.

Mistry, J., A. Bateman, and R.D. Finn. 2007. Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics* **8:** 298.

Monahan, B.J., J.A. Fraser, M.J. Hynes, and M.A. Davis. 2002. Isolation and characterization of two ammonium permease genes, meaA and mepA, from Aspergillus nidulans. *Eukaryot Cell* **1:** 85-94.

Montenegro, E., J.L. Barredo, S. Gutierrez, B. Diez, E. Alvarez, and J.F. Martin. 1990. Cloning, characterization of the acyl-CoA:6-amino penicillanic acid acyltransferase gene of Aspergillus nidulans and linkage to the isopenicillin N synthase gene. *Mol Gen Genet* **221:** 322-330.

Moore, B.S. and C. Hertweck. 2002. Biosynthesis and attachment of novel bacterial polyketide synthase starter units. *Nat Prod Rep* **19:** 70-99.

Mootz, H.D., K. Schorgendorfer, and M.A. Marahiel. 2002. Functional characterization of 4'-phosphopantetheinyl transferase genes of bacterial and fungal origin by complementation of Saccharomyces cerevisiae lys5. *FEMS Microbiol Lett* **213:** 51-57.

Morrissey, J.P. and A.E. Osbourn. 1999. Fungal resistance to plant antibiotics as a mechanism of pathogenesis. *Microbiol Mol Biol Rev* **63:** 708-724.

Nei, M. 1967. Modification of linkage intensity by natural selection. *Genetics* **57:** 625-641.

Nei, M. 2003. Genome evolution: let's stick together. *Heredity* **90:** 411-412.

Nierman, W.C., A. Pain, M.J. Anderson, J.R. Wortman, H.S. Kim, J. Arroyo, M.

Berriman, K. Abe, D.B. Archer, C. Bermejo et al. 2005. Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. *Nature* **438:** 1151-1156.

Nishida, H. 2006. Detection and characterization of fungal-specific proteins in Saccharomyces cerevisiae. *Biosci Biotechnol Biochem* **70:** 2646-2652.

Nixon, J.E., A. Wang, J. Field, H.G. Morrison, A.G. McArthur, M.L. Sogin, B.J. Loftus, and J. Samuelson. 2002. Evidence for lateral transfer of genes encoding ferredoxins, nitroreductases, NADH oxidase, and alcohol dehydrogenase 3 from anaerobic prokaryotes to Giardia lamblia and Entamoeba histolytica. *Eukaryot Cell* **1:** 181-190.

Oberegger, H., M. Eisendle, M. Schrettl, S. Graessle, and H. Haas. 2003. 4'-phosphopantetheinyl transferase-encoding npgA is essential for siderophore biosynthesis in Aspergillus nidulans. *Curr Genet* **44:** 211-215.

Ohno, S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, New York.

Opal, S.M., L.B. Reller, G. Harrington, and P. Cannady, Jr. 1986. Aspergillus clavatus endocarditis involving a normal aortic valve following coronary artery surgery. *Rev Infect Dis* **8:** 781-785.

Orr, W.C. and W.E. Timberlake. 1982. Clustering of spore-specific genes in Aspergillus nidulans. *Proc Natl Acad Sci U S A* **79:** 5976-5980.

Ortiz, D.F., T. Ruscitti, K.F. McCue, and D.W. Ow. 1995. Transport of metal-binding peptides by HMT1, a fission yeast ABC-type vacuolar membrane protein. *J Biol Chem* **270:** 4721-4728.

Osbourn, A.E. 2003. Saponins in cereals. *Phytochemistry* **62:** 1-4.

Osbourn, A.E., B.R. Clarke, P. Lunness, P.R. Scott, and M.J. Daniels. 1994. An oat species lacking avenacin is susceptible to infection by Gaeumannomyces graminis var. tritici. *Physiol. Mol. Plant Pathol.* **45:** 457-467.

Otani, H. 2000. Host Recognition by Plant Pathogens and Role of Host-specific Toxins. *J. Gen. Plant Pathol* **66:** 278-280.

Pain, A., J. Woodward, M.A. Quail, M.J. Anderson, R. Clark, M. Collins, N. Fosker, A. Fraser, D. Harris, N. Larke et al. 2004. Insight into the genome of Aspergillus fumigatus: analysis of a 922 kb region encompassing the nitrate assimilation gene cluster. *Fungal Genet Biol* **41:** 443-453.

Pal, C. and L.D. Hurst. 2003. Evidence for co-evolution of gene order and

recombination rate. *Nat Genet* **33:** 392-395.

Pal, C. and L.D. Hurst. 2004. Evidence against the selfish operon theory. *Trends Genet* **20:** 232-234.

Paoletti, M., C. Rydholm, E.U. Schwier, M.J. Anderson, G. Szakacs, F. Lutzoni, J.P. Debeaupuis, J.P. Latge, D.W. Denning, and P.S. Dyer. 2005. Evidence for sexuality in the opportunistic fungal pathogen Aspergillus fumigatus. *Curr Biol* **15:** 1242-1248.

Paoletti, M., S.J. Saupe, and C. Clave. 2007. Genesis of a fungal non-self recognition repertoire. *PLoS ONE* **2:** e283.

Pasqualotto, A.C. and D.W. Denning. 2006. Post-operative aspergillosis. *Clin Microbiol Infect* **12:** 1060-1076.

Patel, V.B., M. Schweizer, C.C. Dykstra, S.R. Kushner, and N.H. Giles. 1981. Genetic organization and transcriptional regulation in the qa gene cluster of Neurospora crassa. *Proc Natl Acad Sci U S A* **78:** 5783-5787.

Paterson, A.H., J.E. Bowers, M.D. Burow, X. Draye, C.G. Elsik, C.X. Jiang, C.S. Katsar, T.H. Lan, Y.R. Lin, R. Ming et al. 2000. Comparative genomics of plant chromosomes. *Plant Cell* **12:** 1523-1540.

Patron, N.J., R.F. Waller, A.J. Cozijnsen, D.C. Straney, D.M. Gardiner, W.C. Nierman, and B.J. Howlett. 2007. Origin and distribution of epipolythiodioxopiperazine (ETP) gene clusters in filamentous ascomycetes. *BMC Evol Biol* **7:** 174.

Payne, G.A., W.C. Nierman, J.R. Wortman, B.L. Pritchard, D. Brown, R.A. Dean, D. Bhatnagar, T.E. Cleveland, M. Machida, and J. Yu. 2006. *Whole genome comparison of Aspergillus flavus and A. oryzae*. Med Mycol.

Pedras, M.S., L. Irina Zaharia, and D.E. Ward. 2002. The destruxins: synthesis, biosynthesis, biotransformation, and biological activity. *Phytochemistry* **59:** 579-596.

Pel, H.J., J.H. de Winde, D.B. Archer, P.S. Dyer, G. Hofmann, P.J. Schaap, G. Turner, R.P. de Vries, R. Albang, K. Albermann et al. 2007. Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88. *Nat Biotechnol* **25:** 221-231.

Perpetua, N.S., Y. Kubo, N. Yasuda, Y. Takano, and I. Furusawa. 1996. Cloning and characterization of a melanin biosynthetic THR1 reductase gene essential for appressorial penetration of Colletotrichum lagenarium. *Mol*

Plant Microbe Interact **9:** 323-329.

Perrin, R.M., N.D. Fedorova, J.W. Bok, R.A. Cramer, J.R. Wortman, H.S. Kim, W.C. Nierman, and N.P. Keller. 2007a. Transcriptional regulation of chemical diversity in *Aspergillus fumigatus* by *LaeA*. *PLoS Pathog* **3:** e50.

Perrin, R.M., N.D. Fedorova, J.W. Bok, R.A. Cramer, J.R. Wortman, H.S. Kim, W.C. Nierman, and N.P. Keller. 2007b. Transcriptional regulation of chemical diversity in Aspergillus fumigatus by LaeA. *PLoS Pathog* **3:** e50.

Placinta, C.M., J.P.F. D'Mello, and A.M.C. Macdonald. 1999. A review of worldwide contamination of cereal grains and animal feed with Fusarium mycotoxins. *Animal Feed Science and Technology* **78:** 21-37.

Ponticelli, C. 2005. Cyclosporine: from renal transplantation to autoimmune diseases. *Ann N Y Acad Sci* **1051:** 551-558.

Poulton, J.E. 1990. Cyanogenesis in Plants. *Plant Physiol* **94:** 401-405.

Pringle, A., D.M. Baker, J.L. Platt, J.P. Wares, J.P. Latge, and J.W. Taylor. 2005. Cryptic speciation in the cosmopolitan and clonal human pathogenic fungus Aspergillus fumigatus. *Evolution Int J Org Evolution* **59:** 1886-1899.

Proctor, R.H., D.W. Brown, R.D. Plattner, and A.E. Desjardins. 2003. Co-expression of 15 contiguous genes delineates a fumonisin biosynthetic gene cluster in Gibberella moniliformis. *Fungal Genet Biol* **38:** 237-249.

Proctor, R.H., T.M. Hohn, and S.P. McCormick. 1995. Reduced virulence of Gibberella zeae caused by disruption of a trichothecene toxin biosynthetic gene. *Mol Plant Microbe Interact* **8:** 593-601.

Proctor, R.H., R.D. Plattner, D.W. Brown, J.A. Seo, and Y.W. Lee. 2004. Discontinuous distribution of fumonisin biosynthetic genes in the Gibberella fujikuroi species complex. *Mycol Res* **108:** 815-822.

Pryde, F.E., H.C. Gorham, and E.J. Louis. 1997. Chromosome ends: all the same under their caps. *Curr Opin Genet Dev* **7:** 822-828.

Qi, X., S. Bakht, M. Leggett, C. Maxwell, R. Melton, and A. Osbourn. 2004. A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proc Natl Acad Sci U S A* **101:** 8233-8238.

Quiros, L.M., I. Aguirrezabalaga, C. Olano, C. Mendez, and J.A. Salas. 1998. Two glycosyltransferases and a glycosidase are involved in oleandomycin

modification during its biosynthesis by Streptomyces antibioticus. *Mol Microbiol* **28:** 1177-1185.

Rankin, N.E. 1953. Disseminated aspergillosis and moniliasis associated with agranulocytosis and antibiotic therapy. *Br Med J* **1:** 918-919.

Reeves, E.P., K. Reiber, C. Neville, O. Scheibner, K. Kavanagh, and S. Doyle. 2006. A nonribosomal peptide synthetase (Pes1) confers protection against oxidative stress in Aspergillus fumigatus. *Febs J* **273:** 3038-3053.

Rehmeyer, C., W. Li, M. Kusaba, Y.S. Kim, D. Brown, C. Staben, R. Dean, and M. Farman. 2006. Organization of chromosome ends in the rice blast fungus, Magnaporthe oryzae. *Nucleic Acids Res* **34:** 4685-4701.

Reiss, J. 1982. Development of Aspergillus parasiticus and formation of aflatoxin B1 under the influence of conidiogenesis affecting compounds. *Arch Microbiol* **133:** 236-238.

Ricard, G., N.R. McEwan, B.E. Dutilh, J.P. Jouany, D. Macheboeuf, M. Mitsumori, F.M. McIntosh, T. Michalowski, T. Nagamine, N. Nelson et al. 2006. Horizontal gene transfer from Bacteria to rumen Ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics* **7:** 22.

Rice, D.W. and J.D. Palmer. 2006. An exceptional horizontal gene transfer in plastids: gene replacement by a distant bacterial paralog and evidence that haptophyte and cryptophyte plastids are sisters. *BMC Biol* **4:** 31.

Richards, T.A., J.B. Dacks, J.M. Jenkinson, C.R. Thornton, and N.J. Talbot. 2006. Evolution of filamentous plant pathogens: gene exchange across eukaryotic kingdoms. *Curr Biol* **16:** 1857-1864.

Richardson, M.D., S. Rennie, I. Marshall, M.G. Morgan, J.A. Murphy, G.S. Shankland, W.H. Watson, and R.L. Soutar. 2000. Fungal surveillance of an open haematology ward. *J Hosp Infect* **45:** 288-292.

Robbertse, B., J.B. Reeves, C.L. Schoch, and J.W. Spatafora. 2006. A phylogenomic analysis of the Ascomycota. *Fungal Genet Biol* **43:** 715-725.

Robson, G.D., M.G. Wiebe, B. Cunliffe, and A.P. Trinci. 1995. Choline- and acetylcholine-induced changes in the morphology of Fusarium graminearum: evidence for the involvement of the choline transport system and acetylcholinesterase. *Microbiology* **141:** 1309-1314.

Rogers, M.B., N.J. Patron, and P.J. Keeling. 2007. Horizontal transfer of a eukaryotic plastid-targeted protein gene to cyanobacteria. *BMC Biol* **5:** 26.

Rohmer, M., C. Grosdemange-Billiard, M. Seemann, and D. Tritsch. 2004. Isoprenoid biosynthesis as a novel target for antibacterial and antiparasitic drugs. *Curr Opin Investig Drugs* **5:** 154-162.

Rosewich, U.L. and H.C. Kistler. 2000. Role Of Horizontal Gene Transfer In The Evolution Of Fungi. *Annu Rev Phytopathol* **38:** 325-363.

Roy, P.J., J.M. Stuart, J. Lund, and S.K. Kim. 2002. Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. *Nature* **418:** 975-979.

Ryan, F.J., G.W. Beadle, and E.L. Tatum. 1943. The tube method for measuring the growth rate of Neurospora. *Amer. J. Bot.* **30:** 784-799.

Rydholm, C., G. Szakacs, and F. Lutzoni. 2006. Low genetic variation and no detectable population structure in aspergillus fumigatus compared to closely related Neosartorya species. *Eukaryot Cell* **5:** 650-657.

Saeed, A.I., N.K. Bhagabati, J.C. Braisted, W. Liang, V. Sharov, E.A. Howe, J. Li, M. Thiagarajan, J.A. White, and J. Quackenbush. 2006. TM4 microarray software suite. *Methods Enzymol* **411:** 134-193.

Sambrook, J.G., R.D. Campbell, and G. Elgar. 2003. Characterisation of a gene cluster in Fugu rubripes containing the complement component C4 gene. *Gene* **312:** 73-83.

Schardl, C.L., D.G. Panaccione, and P. Tudzynski. 2006. Ergot alkaloids--biology and molecular biology. *Alkaloids Chem Biol* **63:** 45-86.

Scheffer, R.P., R.R. Nelson, and A.J. Ullstrup. 1967. lnheritance of toxin production and pathogenicity in Cochliobolus carbonum and Cochliobolus victoriae. **57:** 1288-1291.

Schrettl, M., E. Bignell, C. Kragl, Y. Sabiha, O. Loss, M. Eisendle, A. Wallner, H.N. Arst, Jr., K. Haynes, and H. Haas. 2007. Distinct roles for intra- and extracellular siderophores during Aspergillus fumigatus infection. *PLoS Pathog* **3:** 1195-1207.

Schwarzer, D. and M.A. Marahiel. 2001. Multimodular biocatalysts for natural product assembly. *Naturwissenschaften* **88:** 93-101.

Schwecke, T., K. Gottling, P. Durek, I. Duenas, N.F. Kaufer, S. Zock-Emmenthal, E. Staub, T. Neuhof, R. Dieckmann, and H. von Dohren. 2006. Nonribosomal peptide synthesis in Schizosaccharomyces pombe and the

architectures of ferrichrome-type siderophore synthetases in fungi. *Chembiochem* **7:** 612-622.

Schweizer, M., M.E. Case, C.C. Dykstra, N.H. Giles, and S.R. Kushner. 1981. Cloning the quinic acid (aq) gene cluster from Neurospora crassa: identification of recombinant plasmids containing both qa-2+ and qa-3+. *Gene* **14:** 23-32.

Schwienbacher, M., L. Israel, J. Heesemann, and F. Ebel. 2005. Asp f6, an Aspergillus allergen specifically recognized by IgE from patients with allergic bronchopulmonary aspergillosis, is differentially expressed during germination. *Allergy* **60:** 1430-1435.

Semon, M. and L. Duret. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol Biol Evol* **23:** 1715-1723.

Seshime, Y., P.R. Juvvadi, I. Fujii, and K. Kitamoto. 2005. Discovery of a novel superfamily of type III polyketide synthases in Aspergillus oryzae. *Biochem Biophys Res Commun* **331:** 253-260.

Shen, B. 2003. Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Curr Opin Chem Biol* **7:** 285-295.

Shimura, K., A. Okada, K. Okada, Y. Jikumaru, K.W. Ko, T. Toyomasu, T. Sassa, M. Hasegawa, O. Kodama, N. Shibuya et al. 2007. Idntification of a biosynthetic gene cluster in rice for momilactones. *J Biol Chem.*

Simillion, C., K. Vandepoele, M.C. Van Montagu, M. Zabeau, and Y. Van de Peer. 2002. The hidden duplication past of Arabidopsis thaliana. *Proc Natl Acad Sci U S A* **99:** 13627-13632.

Sims, J.W., J.P. Fillmore, D.D. Warner, and E.W. Schmidt. 2005. Equisetin biosynthesis in Fusarium heterosporum. *Chem Commun (Camb)*: 186-188.

Singer, G.A., A.T. Lloyd, L.B. Huminiecki, and K.H. Wolfe. 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol Biol Evol* **22:** 767-775.

Smith, M.W., D.F. Feng, and R.F. Doolittle. 1992. Evolution by acquisition: the case for horizontal gene transfers. *Trends Biochem Sci* **17:** 489-493.

Somerville, R.L. and C. Yanofsky. 1964. On The Translation Of The A Gene Region Of Tryptophan Messenger Rna. *J Mol Biol* **8:** 616-619.

Song, Z., R.J. Cox, C.M. Lazarus, and T.T. Simpson. 2004. Fusarin C biosynthesis in Fusarium moniliforme and Fusarium venenatum.

Chembiochem **5:** 1196-1203.

Spellman, P.T. and G.M. Rubin. 2002. Evidence for large domains of similarly expressed genes in the Drosophila genome. *J Biol* **1:** 5.

Squire, R.A. 1981. Ranking animal carcinogens: a proposed regulatory approach. *Science* **214:** 877-880.

Stachelhaus, T., H.D. Mootz, V. Bergendahl, and M.A. Marahiel. 1998. Peptide bond formation in nonribosomal peptide biosynthesis. Catalytic role of the condensation domain. *J Biol Chem* **273:** 22773-22781.

Steenbergen, J.N. and A. Casadevall. 2003. The origin and maintenance of virulence for the human pathogenic fungus Cryptococcus neoformans. *Microbes Infect* **5:** 667-675.

Summerbell, R.C., L. de Repentigny, C. Chartrand, and G. St Germain. 1992. Graft-related endocarditis caused by Neosartorya fischeri var. spinosa. *J Clin Microbiol* **30:** 1580-1582.

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

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673-4680.

Thon, M.R., H. Pan, S. Diener, J. Papalas, A. Taro, T.K. Mitchell, and R.A. Dean. 2006. The role of transposable element clusters in genome evolution and loss of synteny in the rice blast fungus Magnaporthe oryzae. *Genome Biol* **7:** R16.

Tominaga, M., Y.H. Lee, R. Hayashi, Y. Suzuki, O. Yamada, K. Sakamoto, K. Gotoh, and O. Akita. 2006. Molecular analysis of an inactive aflatoxin biosynthesis gene cluster in Aspergillus oryzae RIB strains. *Appl Environ Microbiol* **72:** 484-490.

Townsend, C.A. 1997. Structural studies of natural product biosynthetic proteins. *Chem Biol* **4:** 721-730.

Trail, F., N. Mahanti, M. Rarick, R. Mehigh, S.H. Liang, R. Zhou, and J.E. Linz. 1995. Physical and transcriptional map of an aflatoxin gene cluster in Aspergillus parasiticus and functional disruption of a gene involved early in the aflatoxin pathway. *Appl Environ Microbiol* **61:** 2665-2673.

Tsai, H.F., M.H. Wheeler, Y.C. Chang, and K.J. Kwon-Chung. 1999. A developmentally regulated gene cluster involved in conidial pigment biosynthesis in Aspergillus fumigatus. *J Bacteriol* **181:** 6469-6477.

Tsurushima, T., L.D. Don, K. Kawashima, J. Murakami, H. Nakayashiki, Y. Tosa, and S. Mayama. 2005. Pyrichalasin H production and pathogenicity of *Digitaria*-specific isolates of *Pyricularia grisea*. *Mol. Plant Pathol.* **6:** 605–613.

Tudzynski, P., K. Holter, T. Correia, C. Arntz, N. Grammel, and U. Keller. 1999. Evidence for an ergot alkaloid gene cluster in Claviceps purpurea. *Mol Gen Genet* **261:** 133-141.

Unsold, I.A. and S.M. Li. 2006. Reverse prenyltransferase in the biosynthesis of fumigaclavine C in Aspergillus fumigatus: gene expression, purification, and characterization of fumigaclavine C synthase FGAPT1. *Chembiochem* **7:** 158-164.

Van de Peer, Y. 2004. Computational approaches to unveiling ancient genome duplications. *Nat Rev Genet* **5:** 752-763.

Van de Peer, Y., J.S. Taylor, I. Braasch, and A. Meyer. 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J Mol Evol* **53:** 436-446.

Versteeg, R., B.D. van Schaik, M.F. van Batenburg, M. Roos, R. Monajemi, H. Caron, H.J. Bussemaker, and A.H. van Kampen. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res* **13:** 1998-2004.

Vienken, K. and R. Fischer. 2006. The Zn(II)2Cys6 putative transcription factor NosA controls fruiting body formation in Aspergillus nidulans. *Mol Microbiol* **61:** 544-554.

Vienken, K., M. Scherer, and R. Fischer. 2005. The Zn(II)2Cys6 putative Aspergillus nidulans transcription factor repressor of sexual development inhibits sexual development under low-carbon conditions and in submersed culture. *Genetics* **169:** 619-630.

Villalba, F., J. Collemare, P. Landraud, K. Lambou, V. Brozek, B. Cirer, D. Morin, C. Bruel, R. Beffa, and M.H. Lebrun. 2008. Improved gene targeting in Magnaporthe grisea by inactivation of MgKU80 required for

non-homologous end joining. *Fungal Genet Biol* **45:** 68-75.

Waalwijk, C., T. van der Lee, I. de Vries, T. Hesselink, J. Arts, and G.H.J. Kema. 2004. Synteny in toxigenic Fusarium species: The fumonisin gene cluster and the mating type region as examples. *Eur J Plant Pathol* **110:** 533-544.

Walton, J.D. 2000. Horizontal gene transfer and the evolution of secondary metabolite gene clusters in fungi: an hypothesis. *Fungal Genet Biol* **30:** 167-171.

Watanabe, K., A.P. Praseuth, and C.C. Wang. 2007. A comprehensive and engaging overview of the type III family of polyketide synthases. *Curr Opin Chem Biol* **11:** 279-286.

Weidner, G., C. d'Enfert, A. Koch, P.C. Mol, and A.A. Brakhage. 1998. Development of a homologous transformation system for the human pathogenic fungus Aspergillus fumigatus based on the pyrG gene encoding orotidine 5'-monophosphate decarboxylase. *Curr Genet* **33:** 378-385.

Weigel, B.J., S.G. Burgett, V.J. Chen, P.L. Skatrud, C.A. Frolik, S.W. Queener, and T.D. Ingolia. 1988. Cloning and expression in Escherichia coli of isopenicillin N synthetase genes from Streptomyces lipmanii and Aspergillus nidulans. *J Bacteriol* **170:** 3817-3826.

Wenzl, P., L. Wong, K. Kwang-won, and R.A. Jefferson. 2005. A functional screen identifies lateral transfer of beta-glucuronidase (gus) from bacteria to fungi. *Mol Biol Evol* **22:** 308-316.

Williams, E.J. and D.J. Bowles. 2004. Coexpression of neighboring genes in the genome of Arabidopsis thaliana. *Genome Res* **14:** 1060-1067.

Williams, E.J. and L.D. Hurst. 2002. Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes. *J Mol Evol* **54:** 511-518.

Wolf, J.C. and C.J. Mirocha. 1973. Regulation of sexual reproduction in Gibberella zeae (Fusarium roxeum "graminearum") by F-2 (Zearalenone). *Can J Microbiol* **19:** 725-734.

Wolfe, K.H. and D.C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387:** 708-713.

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

Woolfit, M., E. Rozpedowska, J. Piskur, and K.H. Wolfe. 2007. Genome survey sequencing of the wine spoilage yeast Dekkera (Brettanomyces) bruxellensis. *Eukaryot Cell* **6:** 721-733.

Yabe, K. and H. Nakajima. 2004. Enzyme reactions and genes in aflatoxin biosynthesis. *Appl Microbiol Biotechnol* **64:** 745-755.

Yang, Z. 1997a. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13:** 555-556.

Yang, Z. 1997b. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13:** 555-556.

Yu, J., P.K. Chang, J.W. Cary, M. Wright, D. Bhatnagar, T.E. Cleveland, G.A. Payne, and J.E. Linz. 1995. Comparative mapping of aflatoxin pathway gene clusters in Aspergillus parasiticus and Aspergillus flavus. *Appl Environ Microbiol* **61:** 2365-2371.

Yu, J., P.K. Chang, K.C. Ehrlich, J.W. Cary, D. Bhatnagar, T.E. Cleveland, G.A. Payne, J.E. Linz, C.P. Woloshuk, and J.W. Bennett. 2004. Clustered pathway genes in aflatoxin biosynthesis. *Appl Environ Microbiol* **70:** 1253-1262.

Zakian, V.A. 1996. Structure, function, and replication of Saccharomyces cerevisiae telomeres. *Annu Rev Genet* **30:** 141-172.

Zhang, S., B.J. Monahan, J.S. Tkacz, and B. Scott. 2004. Indole-diterpene gene cluster from Aspergillus flavus. *Appl Environ Microbiol* **70:** 6875-6883.

Zhang, Y., H. Wilkinson, N. Keller, and D. Tsitsigiannis. 2005. *Secondary metabolite gene clusters*. Marcel Dekker, New York.

Zorio, D.A., N.N. Cheng, T. Blumenthal, and J. Spieth. 1994. Operons as a common form of chromosomal organization in C. elegans. *Nature* **372:** 270-272.