

Accepted Manuscript

The 2008 update of the *Aspergillus nidulans* genome annotation: a community effort

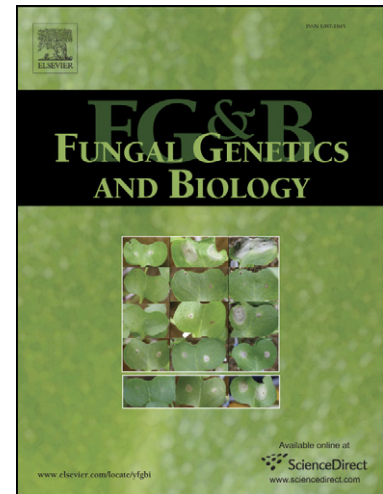
Jennifer Russo Wortman, Jane Mabey Gilsenan, Vinita Joardar, Jennifer Deegan, John Clutterbuck, Mikael R. Andersen, David Archer, Mojca Bencina, Gerhard Braus, Pedro Coutinho, Hans von Döhren, John Doonan, Arnold J.M. Driessen, Pawel Durek, Eduardo Espeso, Erzsébet Fekete, Michel Flippi, Carlos Garcia Estrada, Steven Geysens, Gustavo Goldman, Piet W.J. de Groot, Kim Hansen, Steven D. Harris, Thorsten Heinekamp, Kerstin Helmstaedt, Bernard Henrissat, Gerald Hofmann, Tim Homan, Tetsuya Horio, Hiroyuki Horiuchi, Steve James, Meriel Jones, Levente Karaffa, Zsolt Karányi, Masashi Kato, Nancy Keller, Diane E. Kelly, Jan A.K.W. Kiel, Jung-Mi Kim, Ida J. van der Klei, Frans M. Klis, Andriy Kovalchuk, Nada Kraševc, Christian P. Kubicek, Bo Liu, Andrew MacCabe, Vera Meyer, Pete Mirabito, Márton Miskei, Magdalena Mos, Jonathan Mullins, David R. Nelson, Jens Nielsen, Berl R. Oakley, Stephen A. Osmani, Tiina Pakula, Andrzej Paszewski, Ian Paulsen, Sebastian Pilysyk, István Pócsi, Peter J. Punt, Arthur F.J. Ram, Qinghu Ren, Xavier Robellet, Geoff Robson, Bernhard Seiboth, Piet van Solingen, Thomas Specht, Jibin Sun, Naimeh Taheri-Talesh, Norio Takeshita, Dave Ussery, Patricia A. vanKuyk, Hans Visser, Peter J.I. van de Vondervoort, Ronald P. de Vries, Jonathan Walton, Xin Xiang, Yi Xiong, An Ping Zeng, Bernd W. Brandt, Michael J. Cornell, Cees A.M.J.J. van den Hondel, Jacob Visser, Stephen G. Oliver, Geoffrey Turner

PII: S1087-1845(08)00269-7
DOI: [10.1016/j.fgb.2008.12.003](https://doi.org/10.1016/j.fgb.2008.12.003)
Reference: YFGBI 2084

To appear in: *Fungal Genetics and Biology*

Received Date: 17 October 2008
Revised Date: 15 December 2008
Accepted Date: 15 December 2008

Please cite this article as: Wortman, J.R., Gilsenan, J.M., Joardar, V., Deegan, J., Clutterbuck, J., Andersen, M.R., Archer, D., Bencina, M., Braus, G., Coutinho, P., von Döhren, H., Doonan, J., Driessen, A.J.M., Durek, P., Espeso, E., Fekete, E., Flippi, M., Estrada, C.G., Geysens, S., Goldman, G., de Groot, P.W.J., Hansen, K., Harris, S.D., Heinekamp, T., Helmstaedt, K., Henrissat, B., Hofmann, G., Homan, T., Horio, T., Horiuchi, H., James, S., Jones,



M., Karaffa, L., Karányi, Z., Kato, M., Keller, N., Kelly, D.E., Kiel, J.A.K., Kim, J-M., van der Klei, I.J., Klis, F.M., Kovalchuk, A., Kraševac, N., Kubicek, C.P., Liu, B., MacCabe, A., Meyer, V., Mirabito, P., Miskei, M., Mos, M., Mullins, J., Nelson, D.R., Nielsen, J., Oakley, B.R., Osmani, S.A., Pakula, T., Paszewski, A., Paulsen, I., Pilsyk, S., Pócsi, I., Punt, P.J., Ram, A.F.J., Ren, Q., Robellet, X., Robson, G., Seiboth, B., van Solingen, P., Specht, T., Sun, J., Taheri-Talesh, N., Takeshita, N., Ussery, D., vanKuyk, P.A., Visser, H., van de Vondervoort, P.J.I., de Vries, R.P., Walton, J., Xiang, X., Xiong, Y., Zeng, A.P., Brandt, B.W., Cornell, M.J., van den Hondel, C.A.M., Visser, J., Oliver, S.G., Turner, G., The 2008 update of the *Aspergillus nidulans* genome annotation: a community effort, *Fungal Genetics and Biology* (2008), doi: [10.1016/j.fgb.2008.12.003](https://doi.org/10.1016/j.fgb.2008.12.003)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

The 2008 update of the *Aspergillus nidulans* genome annotation: a community effort

Jennifer Russo Wortman¹, Jane Mabey Gilsean², Vinita Joardar³, Jennifer Deegan⁴, John Clutterbuck⁵, Mikael R. Andersen⁶, David Archer⁷, Mojca Bencina⁸, Gerhard Braus⁹, Pedro Coutinho¹⁰, Hans von Döhren¹¹, John Doonan¹², Arnold J.M. Driessen^{13,14}, Pawel Durek¹¹, Eduardo Espeso¹⁵, Erzsébet Fekete¹⁶, Michel Flippi¹⁷, Carlos Garcia Estrada¹⁸, Steven Geysens¹⁹, Gustavo Goldman²⁰, Piet W.J. de Groot²¹, Kim Hansen²², Steven D. Harris²³, Thorsten Heinekamp²⁴, Kerstin Helmstaedt⁹, Bernard Henrissat¹⁰, Gerald Hofmann^{6,22}, Tim Homan²⁵, Tetsuya Horio²⁶, Hiroyuki Horiuchi²⁷, Steve James²⁸, Meriel Jones²⁹, Levente Karaffa¹⁶, Zsolt Karányi³⁰, Masashi Kato³¹, Nancy Keller³², Diane E. Kelly³³, Jan A.K.W. Kiel^{34,14}, Jung-Mi Kim³⁵, Ida J. van der Klei^{34,14}, Frans M. Klis²¹, Andriy Kovalchuk^{13,14}, Nada Kraševc⁸, Christian P. Kubicek³⁶, Bo Liu³⁵, Andrew MacCabe¹⁷, Vera Meyer^{25,14}, Pete Mirabito³⁷, Márton Miskei³⁸, Magdalena Mos²⁹, Jonathan Mullins³³, David R. Nelson³⁹, Jens Nielsen^{6,40}, Berl R. Oakley²⁶, Stephen A. Osmani⁴¹, Tiina Pakula⁴², Andrzej Paszewski⁴³, Ian Paulsen⁴⁴, Sebastian Pilzyk⁴³, István Pócsi³⁸, Peter J. Punt⁴⁵, Arthur F.J. Ram^{25,14}, Qinghu Ren³, Xavier Robellet⁴⁶, Geoff Robson⁴⁷, Bernhard Seiboth³⁶, Piet van Solingen⁴⁸, Thomas Specht⁴⁹, Jibin Sun⁵⁰, Naimeh Taheri-Talesh⁹, Norio Takeshita⁵¹, Dave Ussery⁵², Patricia A. vanKuyk²⁵, Hans Visser⁵³, Peter J.I. van de Vondervoort⁵⁴, Ronald P. de Vries⁵⁵, Jonathan Walton⁵⁶, Xin Xiang⁵⁷, Yi Xiong⁴¹, An Ping Zeng⁵⁸, Bernd W. Brandt⁵⁹, Michael J. Cornell^{46,60}, Cees A.M.J.J. van den Hondel^{25,14}, Jacob Visser^{25,61}, Stephen G. Oliver⁶² and Geoffrey Turner⁶³

¹Department of Medicine Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA. ²School of Medicine, University of Manchester, Manchester, UK. ³The J. Craig Venter Institute, Rockville, MD, USA. ⁴European Bioinformatics Institute, Hinxton, Cambridge, UK. ⁵Faculty of Biomedical and Life Sciences, University of Glasgow, Glasgow, Scotland, UK. ⁶Center for Microbial Biotechnology, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark. ⁷School of Biology, University of Nottingham, Nottingham, UK. ⁸National Institute of Chemistry, Laboratory for Biotechnology and Industrial Mycology, Ljubljana, Slovenia. ⁹Institut für Mikrobiologie & Genetik, Georg-August-Universität, Göttingen, Germany. ¹⁰Université de Provence (Aix- Marseille I), Marseille, France. ¹¹Technical University Berlin, Berlin, Germany. ¹²Department of Cellular and Developmental Biology, John Innes Centre, Norwich, UK. ¹³Department of Molecular Microbiology, Groningen Biomolecular Sciences and Biotechnology Institute, Haren, The Netherlands. ¹⁴Kluyver Centre for Genomics of Industrial Microorganisms, The Netherlands. ¹⁵CIB-CSIC, Madrid, Spain. ¹⁶Department of Genetics and Applied Microbiology, Faculty of Science, University of Debrecen, Debrecen, Hungary. ¹⁷Instituto de Agroquímica y Tecnología de Alimentos, CSIC, Valencia, Spain. ¹⁸Instituto de Biotecnología de León (INBIOTEC), León Spain. ¹⁹VIB Department for Molecular Biomedical Research, Gent, Belgium. ²⁰Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Brazil. ²¹Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands. ²²Novozymes A/S, Bagsvaerd, Denmark. ²³Plant Science Initiative, University of Nebraska, Lincoln, NE, USA. ²⁴Department of Molecular and Applied Microbiology, Leibniz-Institute for Natural Product Research and Infection Biology (HKI) Jena, Germany. ²⁵Section Molecular Microbiology, Institute of Biology, Leiden University, Leiden, The Netherlands. ²⁶Department of Molecular Biosciences, The University of Kansas, Lawrence, KS, USA. ²⁷Department of Biotechnology, The University of Tokyo, Japan. ²⁸Gettysburg College, Department of Biology, Gettysburg, PA, USA. ²⁹School of Biological Sciences, University of Liverpool, Liverpool, UK. ³⁰First Department of Medicine, Faculty of Medicine, University of Debrecen, Debrecen, Hungary. ³¹Dept Biological Mechanisms and Functions, Graduate School of Bioagricultural Sciences, Furo-cho, Chikusa-ku, Nagoya, Japan. ³²Dept of Plant Pathology, Univ. of Wisconsin, Madison, WI, USA. ³³Institute of Life Sciences, The School of Medicine, Swansea University, Swansea, UK. ³⁴Molecular Cell Biology, Groningen Biomolecular Sciences and Biotechnology Institute (GBB), University of Groningen, Haren, The Netherlands. ³⁵Section of Plant Biology, University of California, Davis, CA, USA. ³⁶Vienna University of Technology, Institute of Chemical Engineering, Division of Gene Technology and Applied Biochemistry, Vienna, Austria. ³⁷University of Kentucky, School of Biological Sciences, Lexington, KY, USA. ³⁸Department of Microbial Biotechnology and Cell Biology, Faculty of Science and Technology, University of Debrecen, Debrecen, Hungary. ³⁹Department of Molecular Sciences, University of Tennessee, Memphis, USA. ⁴⁰Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden.

⁴¹Department of Molecular Genetics, Ohio State University, Columbus, OH, USA.

⁴²VTT Technical Research Centre of Finland, Espoo, Finland. ⁴³Institute of Biochemistry and Biophysics, Department of Genetics, Warsaw, Poland.

⁴⁴Department of Chemistry and Biomolecular Sciences, Macquarie University

Sydney, Australia. ⁴⁵TNO Quality of Life, Industrial Biotechnology, Zeist, The

Netherlands. ⁴⁶Institut de Génétique et Microbiologie, CNRS - Université Paris-Sud

XI, Orsay, France. ⁴⁷Faculty of Life Sciences, University of Manchester, Manchester,

UK. ⁴⁸Danisco-Genencor, Leiden, The Netherlands. ⁴⁹Sandoz GmbH, Kundl, Tirol,

Austria. ⁵⁰Helmholtz Centre for Infection Research, Braunschweig, Germany.

⁵¹University of Karlsruhe, Department of Applied Biosciences Applied Microbiology,

Karlsruhe, Germany. ⁵²Center for Biological Sequence Analysis BioCentrum-DTU, The

Technical University of Denmark, Lyngby, Denmark. ⁵³Dyadic Nederland BV,

Wageningen, The Netherlands. ⁵⁴DSM Anti-Infectives, Delft, The Netherlands.

⁵⁵Microbiology, Utrecht University, Utrecht, The Netherlands. ⁵⁶DOE Plant Research

Lab, Michigan State University, East Lansing, MI, USA. ⁵⁷Department of Biochemistry,

USUHS, Bethesda, MD, USA. ⁵⁸Hamburg University of Technology, Institute of

Bioprocess and Biosystems Engineering, Hamburg, Germany and Helmholtz Centre

for Infection Research, Braunschweig, Germany. ⁵⁹Centre for Integrative

Bioinformatics (IBIVU), VU University of Amsterdam, The Netherlands. ⁶⁰School of

Computer Science, University of Manchester, Manchester, UK. ⁶¹Fungal Genetics

and Technology consultancy, Wageningen, The Netherlands. ⁶²Department of

Biochemistry, University of Cambridge, Cambridge, UK. ⁶³Department of Molecular

Biology and Biotechnology, University of Sheffield, Sheffield, UK. Correspondence

should be addressed to G.T. (g.turner@sheffield.ac.uk)

Abstract

The identification and annotation of protein-coding genes is one of the primary goals of whole-genome sequencing projects, and the accuracy of predicting the primary protein products of gene expression is vital to the interpretation of the available data and the design of downstream functional applications. Nevertheless, the comprehensive annotation of eukaryotic genomes remains a considerable challenge. Many genomes submitted to public databases, including those of major model organisms, contain significant numbers of wrong and incomplete gene predictions. We present a community-based reannotation of the *Aspergillus nidulans* genome with the primary goal of increasing the number and quality of protein functional assignments through the careful review of experts in the field of fungal biology.

Index descriptors: *Aspergillus nidulans*; aspergilli; genome; annotation; fungal community; assembly; transcription factors; CADRE

1) Introduction

Importance of *A.nidulans* as a reference for the eight sequenced aspergilli

Genome sequences of 8 species of the genus *Aspergillus* have now been determined (Table 1). However, the aspergillus research community is relatively small, and the task of detailed annotation of all of these genomes, together with presentation of the information in an easily accessible form, is enormous. We have therefore chosen to begin the task with the genetic model species *Aspergillus nidulans*.

i) General importance of aspergilli for fungal biology, human health, biotechnology, and agriculture

The genus *Aspergillus* comprises a diverse group of filamentous fungi. Despite belonging to the same genus, *Aspergillus* species have diverged significantly (Galagan et al., 2005), though they are sufficiently related such that orthologues can be identified for the majority of genes. *A. nidulans*, the most studied species, has been an important model organism for eukaryotic genetics for over 60 years (Martinelli and Kinghorn, 1994), with the advantage of having a sexual cycle (teleomorph *Emericella nidulans*), which is usually absent from other species of *Aspergillus*. In addition to a long history of classical genetic and biochemical studies, most molecular techniques for gene manipulation were first developed in *A. nidulans* before application to other members of the genus. Detailed laboratory protocols have recently been made easily accessible (Osmani et al., 2006; Szewczyk et al., 2006; Todd et al., 2007a; Todd et al., 2007b), and mutant strains isolated over many years are available from the Fungal Genetics Stock Centre (www.fgsc.net), together with other useful resources such as vectors and libraries.

It is important to note that different wild-type strains of *A. nidulans* exist (Jinks et al., 1966), but all the commonly used mutant strains are derived from a single strain, sometimes called

the Glasgow strain, following its choice as a genetic model in the early 1950s (Pontecorvo et al., 1953). Our detailed understanding of the genetics and physiology of *A. nidulans* provides an excellent basis for extension of this knowledge to other, imperfect species of economic importance. These include the opportunistic pathogen *A. fumigatus*, a cause of allergies and a growing threat to immunocompromised patients; *A. niger* and *A. oryzae*, sources of industrial enzymes and other products such as citric acid; *A. flavus*, a plant pathogen and toxin-producing agricultural spoilage organism; *A. terreus*, sometimes an opportunistic pathogen, is also a source of lovastatin, one of the first of the hugely successful statins used therapeutically to inhibit cholesterol biosynthesis.

A. nidulans possesses a penicillin biosynthesis pathway similar to that found in the industrial producer *Penicillium chrysogenum*, and has most of the steps of the aflatoxin pathway found in *A. flavus* and *A. parasiticus*, so that it has become a key model system for studying these secondary metabolic pathways and their regulation (Bok et al., 2006b; Brakhage et al., 2004).

A. nidulans genetics has also contributed to eukaryotic cell biology beyond its close fungal relatives, examples being the discovery of γ -tubulin (Oakley and Oakley, 1989) and NudF, a homologue of the human lysencephaly protein (Xiang et al., 1995).

Since *A. nidulans* has played such an important role as a genetic model, Eurofung through the Eurofungbase Project decided to focus its community annotation efforts on *A. nidulans* in the first instance. We report the major findings of this exercise here, specific accounts on particular biological domains are presented in the following series of papers, and the newly annotated *A. nidulans* genome sequence is available in the CADRE database (Mabey et al., 2004) and will also be available in the Fungi section of ENSEMBL genomes at the EBI.

ii) Annotation of aspergilli

The genome sequences of eight distinct *Aspergillus* species have been publicly released over the past four years. However, many of these genomes were annotated at different institutions using diverse methods over a relatively long time period, during which available tools and datasets have evolved rapidly. The inconsistency in annotation quality and completeness across these species hinders many avenues of comparative genomic research that depend on high-quality genome annotation, including evolutionary and functional studies (Wortman et al., 2006b).

The first three *Aspergillus* genome sequences, those for *A. fumigatus*, *A. oryzae* and *A. nidulans*, were published in 2005 and described in three companion papers (Galagan et al., 2005; Machida et al., 2005; Nierman et al., 2005). The *A. fumigatus* genome sequence (Af293) was generated through a collaboration between the Institute for Genomic Research (TIGR) and the Wellcome Trust Sanger Institute and deposited in GenBank; it has the accession **AAHF00000000** (Nierman et al., 2005). The assembled genomic sequence was processed through the TIGR annotation pipeline, which subjected each sequence to a series of homology searches as well as algorithms for predicting genes (GlimmerM, Exonomy, Unveil, and GeneSplicer) (Majoros et al., 2003; Pertea et al., 2001). The gene prediction algorithms were trained with a limited dataset of *A. fumigatus* EST and cDNA sequences (Nierman et al., 2005) and the output of the pipeline was manually reviewed. An updated annotation for *A. fumigatus*, based on comparative genome data and involving targeted manual annotation, was released to GenBank in March, 2007 (Fedorova et al., 2008). The *A. oryzae* RIB40 genome, sequenced and analyzed by a consortium led by the National Institute of Advanced Industrial Science and Technology (AIST) in Japan, was released to DDBJ under the accession numbers **AP007150** to **AP007177**. The AIST annotation pipeline incorporated EST and protein homology data via the gene predictors ALN (Gotoh, 2000), GeneDecoder (Asai et al., 1998) and GlimmerM. The pipeline was trained on a set of gene models which were constructed by alignment of known fungal proteins to the open reading-frames in the *A. oryzae* genome (Machida et al., 2005). *A. oryzae* annotation was last updated in DDBJ in December 2005. The *A. nidulans* (*Emericella nidulans*) genome, for strain FGSC A4, was sequenced and annotated by the

Broad Institute of Harvard and MIT and submitted to GenBank and has the accession **AACD00000000**. The genome sequence was annotated using the Calhoun annotation system, which included protein homology searches and the gene prediction algorithms FGENESH (Salamov and Solovyev, 2000), FGENESH+, and GENEWISE (Birney et al., 2004). *A. nidulans* EST data was not incorporated into the gene predictions, but was used separately for validation (Galagan et al., 2005).

Comparison of these first aspergillus genome sequences revealed a surprising level of genetic variability. Proteome comparisons revealed an average amino-acid identity of less than 70% between each species pair, suggesting that they are as evolutionarily distant from each other as humans are from fish (Galagan et al., 2005). Since these phylogenetic distances were so significant, it became clear that additional aspergilli would need to be sequenced to facilitate comparative analyses. More data would be needed to elucidate the specific gene differences and regulatory elements linked to the distinctive phenotypic and physiological properties important to the study of each organism. The proteome comparison also revealed the extent of gene model annotation differences between genomes, with the majority of identified orthologue groups (~80%) containing members differing in length and/or number of exons. Identified annotation problems included the merging of neighbouring loci, missed exon calls, and incorrect 5' exons (Wortman et al., 2006b). A summary of the published genome sequences for *Aspergillus* species is given in Table 1, which provides information on the sequencing centres, software tools employed, and access to the sequences.

In addition to inconsistencies in gene model annotation caused by the lengthy timeframe over which the different genomes were sequenced and the diverse annotation processes employed at different sequencing centres, there are also significant differences in the functional annotation attached to gene products. While some groups will only attach a putative function based on a high stringency homology match to an experimentally characterized protein (Galagan et al., 2005), others have used more lenient criteria, employing resources such as InterPro (Mulder and Apweiler, 2007) and PFAM (Finn et al.,

2008) profiles (Fedorova et al., 2008; Nierman et al., 2005) and the NCBI KOG (Tatusov et al., 2003) resource (Machida et al., 2005).

Additional attributes, such as Enzyme Commission numbers (Bairoch, 2000) or Gene Ontology associations (Ashburner et al., 2000) are also not consistently applied. As part of the original primary annotation efforts, the functional annotation of *A. fumigatus* (Af293) and *A. niger* (CBS 513.88) were manually reviewed with input from the research community, and with an emphasis on particular protein families (Fedorova et al., 2008; Nierman et al., 2005; Pel et al., 2007). Manual annotation by domain experts is a valuable approach for integrating multiple lines of computational evidence, but is a resource-intensive and time-consuming process. Thus, it is best applied to species within a given clade for which there is most genetic and functional data and for which the largest and most active research community exists. For all of these reasons, a community annotation of the *A. nidulans* genome sequence appeared a high priority. Having a well-annotated reference genome for the aspergillus clade of organisms will support the transitive improvement of the annotation across orthologous genes in the other members of the genus.

Eurofungbase

Eurofungbase is a coordination action programme funded by the European Commission under contract LSSG-CT-2005-018964. It comprises a community of 32 different partner laboratories in 11 European countries, supported by an Industry Platform of 13 companies. Eurofungbase aims to facilitate the construction of an integrated data warehouse to enable comparative and functional genomic studies of filamentous fungi of scientific, medical, industrial, and agricultural importance. The Consortium has a number of different strategies for achieving its aims, but one of the most important approaches is to organise the manual annotation of the genomes of important model organisms by expert groups of researchers, and to consolidate and disseminate the results of such annotation exercises through journal publications and web facilities. It is within this context that Eurofungbase took on the task of re-annotating the *Aspergillus nidulans* genome with the help of colleagues from TIGR/J Craig Venter Institute and aspergillus research laboratories in the USA.

Evolution of *A. nidulans* genome data

The original public genome annotation of *A. nidulans*, described briefly above, consists of 9,541 protein-coding gene predictions (Galagan et al., 2005). Each gene was assigned a unique locus identifier with the prefix AN followed by a four digit number between 0001 and 9541 and appended with the annotation version number 2 (e.g. AN0001.2). Version 1 was internal to the Broad Institute and not released widely. As of July 2008, this is the version of the *A. nidulans* annotation that is still represented at GenBank, linked to accession [AACD00000000](#), submitted in January, 2004. Of the genes overlapping EST alignments, approximately 70% were fully consistent with the EST data, while 30% showed some inconsistency (J. E. Galagan, personal communication). Comparative analysis with *A. fumigatus* and *A. oryzae* suggested that there were many neighboring loci inappropriately merged (Wortman et al., 2006b). Functional annotation was applied to gene products only when they exhibited high-identity matches to previously published, experimentally characterised, proteins within the fungal kingdom. This resulted in putative function assignments for approximately 3% of the predicted proteins.

In summer, 2005, NIAID requested that the annotation group at TIGR revisited the gene structure annotation of *A. nidulans* in advance of microarray design being planned by the NIAID-funded Pathogen Functional Genomics Resource Center (PFGR). This re-annotation effort focused on the automated incorporation of EST data into the existing gene models, and the manual review and correction of merged loci. 32,931 EST and cDNA sequences compiled from GenBank and provided by C. d'Enfert and G. Goldman were aligned to the genome, and compared to the existing annotation using the PASA pipeline (Haas et al., 2003). These EST sequences collapsed to 8,690 unique assemblies and were used to perform automated gene structure updates of 1,146 genes. In addition, over 2,000 genes that could not be computationally resolved with the current gene structure were manually reviewed and corrected on the basis of either protein homology or EST data. 494 loci were split into two or more distinct loci, and 214 new gene models were added. In addition, 426 gene models originally predicted by the Broad institute, but excluded from the earlier release because they did not meet minimum length criteria,

were also incorporated. The final gene set consisted of 10,701 protein-coding gene predictions, with 4,263 genes completely consistent with EST alignments. Locus identifiers were retained in all cases of one-to-one mapping (9447), whether the sequence changed or not, but the version number was incremented to 3 for all genes. Since the gene number was now over 10,000, there was a need to create new locus identifiers with 5 numeric digits after the AN prefix (e.g. AN10002.3). Functional annotation was supplied by the Broad Institute. As of July, 2008, this version 3 annotation data set is the current annotation reflected at the Broad Institute web site: http://www.broad.mit.edu/annotation/genome/aspergillus_group/.

2. Eurofungbase Community Annotation

Main focus: Gene function

The primary goal of the Eurofungbase annotation effort is to increase the numbers of *A. nidulans* proteins with informative functional assignments. Experts in various aspects of fungal and, particularly, *A. nidulans* biology were invited to participate in an ongoing annotation effort, which started with a jamboree in Autumn 2006. Prior to this initial meeting, the version 3 protein sequences were subjected to a series of computational analyses intended to provide evidence for protein function. These included homology searches against the GenBank non-redundant database (nr), domain identification against the PFAM(Finn et al., 2008) and InterPro(Mulder and Apweiler, 2007) databanks, and the programs tmHMM(Sonnhammer et al., 1998), which predicted transmembrane domains, and SignalP (Bendtsen et al., 2004), which predicts signal peptides.

To support the annotation effort, meeting participants were granted access to a compact summary of this computational evidence through the Manatee web-based interface (Wortman et al., 2006a). Manatee is a web-based manual annotation and analysis tool that acts as an interface between an underlying database and an annotator. Manatee's interface allows the annotator to add, delete, and edit annotations attached to the protein, such as gene product names, gene symbols, EC numbers and GO assignments. All such changes are referred back to the underlying database. Consortium members could continue to interact with the annotation database, through Manatee, on an ongoing basis for more than a year.

Over the course of the functional annotation effort, 2,626 genes were reviewed and edited, with GO terms being added where possible and appropriate. Through this concerted effort, the percentage of *A. nidulans* gene products with an informative name has increased from approximately 3% to 19%. For the remaining un-reviewed genes, we

have provided product names by transfer of information from *A. fumigatus* or *A. niger* orthologues, further increasing the proportion of gene products with informative names to 58%. As an indication of the changes arising from this project, we have also incremented the annotation version to 4 for all locus identifiers (e.g., AN****.4).

b. Assembly updates

Assembly of supercontigs and association with chromosomes

The initial Broad Institute assembly comprised 173 contigs, linked by end-sequenced BAC and fosmid bridges to form 16 supercontigs, corresponding to the 16 chromosome arms. A further 75 contigs were unplaced. Using BLAST (Altschul et al., 1990) and BL2seq (Tatusova and Madden, 1999) it was found that many of the Broad contigs overlapped so that 58 previously unassigned contigs could be incorporated into supercontigs, and the number of unsequenced gaps within supercontigs was reduced from 157 to 71 (see <http://www.gla.ac.uk/ibls/molgen/aspergillus/contiglinks.html>

and http://www.cadre-genomes.org.uk/Aspergillus_nidulans/Docs/revised_contig_overlaps.html). In five further cases, gaps are bridged by independently cloned sequences or retrotransposons with matching target-site duplications. The remaining gaps are assigned an arbitrary length of 1 kb in the revised assembly.

Supercontigs were related to chromosome arms using meiotically mapped and cloned linkage map markers (see <http://www.gla.ac.uk/ibls/molgen/aspergillus/index.html>). Over 185 such markers, identifiable with auto-called genes, are currently informative in this respect. With a few exceptions, most of which can be put down to reliance on inadequate linkage data, the correspondence between linkage and genome maps is excellent (Clutterbuck and Farman, 2008). Four supercontigs have telomeric simple sequence (TTAGGG) repeats at one end, allowing the orientation of the arm to be validated independently of the genetic map, and using TERMINUS (Li et al., 2005), Clutterbuck and Farman (2008) were able to identify telomeric simple sequences and

subtelomeric contigs in the NCBI sequence trace archive and associate them with the remaining chromosome arms. While six different telomeres have identical telomeric contigs, the subtelomeric sequences are more varied, including a variety of simple sequences, and, in many cases, members of a specific class of telomere-linked helicases, all but one of which are truncated or degenerate (Clutterbuck and Farman, 2008). In most cases, the non-telomeric ends of supercontig arms were marked with clusters of A+T-rich, degraded transposable elements that are typical of centromeric DNA. In the revised assembly, centromeres, which have not been sequenced, are displayed as arbitrary gaps of 50 kb.

The one case in which there was a serious discrepancy between genomic and linkage maps concerned supercontig 6, associated with chromosome V. Here a telomeric link was found in the middle of the supercontig, requiring the splitting of this supercontig into 6a and 6b. New supercontig 6a was found to end in a short fragment of a ribosomal RNA repeat sequence and now spans a central region of chromosome arm V-L, between the nucleolus organizer and the centromere, the distal section of V-L being formed by supercontig 12. Supercontig 6b, starting with links to telomere T15, makes up chromosome arm V-R (see Fig 5.6 in Clutterbuck and Farman, 2008). The resulting rearranged supercontigs now correspond well with the linkage map of chromosome V, already known to include the nucleolus organizer (Brody et al., 1991), (approximate length 360 kb: Ganley and Kobayashi, 2007).

c) Assembly availability

In summary, the new assembly has 248 contigs, of which 231 contigs are assigned to 17 supercontigs (with 66 unsequenced gaps) that are mapped to eight linkage groups (approx. 30.5Mb). This assembly is displayed within the Central *Aspergillus* Data Repository (CADRE) (Mabey et al., 2004). However, due to the recent improvements in assembly and annotation data, the current contigs no longer reflect the annotation within GenBank although the sequence remains the same. For this reason, contigs in CADRE are labeled with source identifiers from the Broad Institute (1.1 to 1.248), which coincide with the public sequence data within GenBank ([AACD01000001](#) to [AACD01000248](#)). The supercontigs, annotated as scaffolds within GenBank ([CH236920](#) to [CH236935](#)) have also changed.

Therefore we have labeled the supercontigs in CADRE 1 to 16, with supercontig 6 replaced by 6a and 6b. These supercontigs are assigned to the linkages groups, which remain labeled as I to VIII.

d) Gene structure updates

The version 3 gene models, which were annotated on the 248 original contig sequences, were mapped to the revised chromosome sequences. The genome sequences were aligned using the NUCmer utility of the MUMmer package (Kurtz et al., 2004), and gene models were transferred on the basis of the resulting coordinate mapping. In this process, 82 gene predictions were unable to be transferred because of changes in the underlying sequence, and approximately 200 mapped gene pairs were found to overlap significantly. 34 gene predictions were marked as putative pseudogenes, as the open reading-frames are disrupted in the reported genome sequence. The transferred gene models were the starting point for version 4 of the *A. nidulans* annotation.

During the course of functional annotation, consortium members with expertise in specific gene families were able to identify gene structures that were likely to be incorrect. They then submitted either virtual cDNA sequences representing the corrected gene structure, or protein sequences corresponding to corrected genes. PASA (Haas et al., 2003) was used to correct genes based on the submitted cDNA sequences, processing 74 gene structure updates, including 7 additional putative pseudogenes. Genewise (Birney et al., 2004) was used to instantiate gene structures based on the submitted protein sequences. These structures were reviewed manually and used to update 95 genes, including 3 additional putative pseudogenes, and to create five new gene models. The final gene count for the version 4 annotation is 10,605, which includes 63 putative pseudogenes. Two families of genes were significantly improved through these efforts. 39 of 119 cytochrome P450 genes were updated, as were 96 of the 342 predicted Zn(II)2Cys6 transcription factors.

This dataset reflects an iterative improvement in the *A. nidulans* predicted gene complement, but should not be considered a final product. The gene structure annotation of *A. nidulans* will continue to change over time as new experimental evidence and computational approaches arise. Of particular interest are recently published gene prediction programs that use machine-learning approaches to improve performance on eukaryotic genomes (Bernal et al., 2007; DeCaprio et al., 2007). The annotation process can also benefit from the incorporation of more diverse data types and predictions. One example is found in the companion article by Wang et al. FGB-08-96, which describes a new and species-specific intron splice site predictor for aspergillus. In contrast to conventional gene finders producing entire gene structures, splice site predictors identify all candidate splice sites and assign probability values, information which can be used to correct gene structures or predict alternatively spliced genes.

New findings from the annotation

Subsequent articles in this issue do not aim to cover all the different families of genes, but rather reflect current research interests of the aspergillus community. Some of the genes included in the annotation (over 800) have been functionally characterised and/or genetically mapped prior to completion of the genome sequence, and these genes are mostly named according to the *A. nidulans* convention (Clutterbuck, 1973; Martinelli, 1994), in addition to the locus identifier in the AN**** format. Where genes have not been functionally characterised, the gene model name based on bioinformatics analysis is given only in the AN**** designation. For these uncharacterised genes, annotation is based on comparative genomics, and functions are suggested prior to functional analysis.

Since a genetic map of the 8 linkage groups (chromosomes) of *A. nidulans* was produced prior to the genome sequencing project, it was instructive to compare the genetic and physical maps. In most cases, the correlation is good, and in the case of chromosome V, the linkage map assisted in the correct assembly of the genome sequence.

a) Transposable elements

Transposable elements (TEs) were catalogued mainly by V.V.Kapitonov & J.Jurka, Genetic Information Research Institute, and can be found in Repbase (<http://girinst.org/repbase/>). 1275 insertions of whole or fragmented elements were identified, representing 14 retroposons and 19 DNA transposon families. Their sequences and distribution are described elsewhere (Clutterbuck et al., 2008).

A total of 306 autocalled genes are involved with TEs identified by Kapitonov & Jurka. 151 are found wholly within TEs and many of these have recognizable transposition-related features. 103 genes overlap TEs or TE fragments, and 52 others have small TEs or fragments within them, only 14 of these being wholly within introns. In many cases the TEs are affected by RIP, and their associated autocalled genes have atypical (probably unrealistic) structures, often including multiple long introns, e.g. the mean length of 67 introns in 24 genes overlapping Mariner-6 elements was 152 bp: three times the modal introns length.

In only two cases have the affected genes been previously identified: the 3' end of AN7818.3 (*stcF*) overlaps a degraded *Mariner-4_AN* copy by just three nucleotides. Secondly, (Cultrone et al., 2007) have reported a case where a non-autonomous *Helitron-N1_AN* element has been created as a result of a 3' deletion of an autonomous *Helitron-1-AN* element and readthrough into the *xanA* gene (AN10081.3). The new element has transposed once to give a second copy of the *xanA* promoter and 5' region, named *psxA* (AN11581.3). This pseudogene is transcribed, but the resulting mRNA is rapidly destroyed by nonsense-mediated decay.

b) Transcription factors

Eukaryotic transcription factors can be classified on the basis of their characteristic DNA-binding domains. The *A. nidulans* genome was analyzed for the presence of twelve different classes of DNA-binding transcription factors (Table 2). DNA-binding proteins without transcription factor activity were not included in the analysis. To identify putative transcription factors in the *A. nidulans* genome, BlastP searches were performed using functionally characterized transcription factors and/or predicted transcription factors from *Saccharomyces cerevisiae* and *Neurospora crassa* as queries. Predicted proteins that showed significant similarity to the above-mentioned transcription factors, but lacked the

typical DNA-binding domain, were subsequently checked manually for alternative gene models that would include a DNA-binding domain. An overview of the number of putative transcription factors in the *A. nidulans* genome is given in Table 2 and further details are given as supplementary data (Supp. Tables 1-13).

As shown in Table 2, the family of Zn(II)2Cys₆ transcription factors (MacPherson et al., 2006) is the largest, consisting of 330 proteins. The number of annotated proteins belonging to this class has been increased significantly from the initial annotation (Galagan et al., 2005) by improving the gene models manually. By addition of 5'exons, and/or changing intron/exon boundaries, in many cases a complete conserved six cysteine motif (CX₂CX₆CX₅₋₁₆CX₂CX₆₋₈C) that was absent in a large number of the original gene models could be included in the protein sequence. For only a small number of putative Zn(II)2Cys₆ TF proteins were we unable to identify a specific DNA binding domain even by manual annotation (Suppl. Table 13)

Members of an interesting subclass of Zn(II)2Cys₆ transcription factors contain a C2H2 DNA binding domain in addition to the Cys₆ motif. A total of 12 such transcription factors have been identified (Supp. Table 2). The function of those genes has not been studied in *A. nidulans*. In *Colletotrichum lagenarium* (Cmr1) and *Magnaporthe griseae* (Pig1) the double motif transcription factors are involved in the regulation of mycelial melanin biosynthesis (Tsuji et al., 2000).

It is a major challenge for future work to perform functional analysis of the members of the different families of transcription factors. High-throughput deletion strategies or overexpression approaches are useful methods for studying their function, as has proved to be the case for *N. crassa* (Colot et al., 2006) For *A. nidulans*, only 21 Zn(II)2Cys₆ transcription factors have been functionally analyzed in more detail (Suppl. Table 1).

Previous work has indicated that, in various *Aspergillus* species, transcription factors are often clustered in the genome with their target genes (Cazelle et al., 1998; Felenbok et al., 2001; Gomi et al., 2000; Hull et al., 1989; Lamb et al., 1990; Unkles et al., 1992; Woloshuk et al., 1994; Yuan et al., 2008a). This applies, for example, to genes associated with primary and secondary metabolism and carbohydrate metabolism. Based on this finding, analysis of genes adjacent to the Zn(II)2Cys₆ transcription factors might give information about their possible target genes and, hence, function (Flipphi et al., 2006; Ram and Punt, unpublished results). In addition, phylogenetic analyses of the protein sequences of

Zn(II)2Cys₆ transcription factors may be useful for the identification of sub-clusters with related functions. In a simple phylogenetic analysis of the 330 *A. nidulans* Zn(II)2Cys₆ transcription factors, approximately 56 sub-clusters (defined as proteins with a reciprocal BlastP e-value of < -20 towards each other) could be identified. One of these clusters contained both the AmyR(AN2016.4) and InuR (AN3835.4) transcription factors involved in the regulation of enzymes degrading starch and inulin, respectively (Yuan et al., 2008a) . The possible roles in polysaccharide catabolism of other transcription factors in this specific subcluster (AN7343.4, AN7346.4, AN8596.4 and AN10423.4) await further analysis.

c) Polysaccharide degrading enzymes

A. nidulans is found naturally in the rhizosphere, where it is able to utilise decaying plant material by means of secreted protein- and polysaccharide-degrading enzymes. Despite significant differences in the set of putative plant polysaccharide-degrading enzymes predicted from the genome sequence of *A. niger*, *A. nidulans*, and *A. oryzae*, no large differences were found in their growth on cellulose, xylan, pectin or galactomannan (Coutinho et al. FGB-08-108). The apparent differences in the repertoire of enzymes may, therefore, reflect specific adaptations to the natural biotope of the individual species. The identification of transcription factor (TF) target sites within the promoter regions of genes encoding polysaccharide-degrading enzymes is rarely helpful in predicting function. Thus while there is a correlation between the presence of XlnR sites and xylose induction for *A. niger* genes, this is not strictly true for the other two species; it is likely that TF binding sites in the promoters vary between *Aspergillus* species.

d) Primary metabolism

Primary metabolism and its regulation has been a major pre-genomic research area, and the post-genomic annotation has now identified genes for most of the central pathways. Several of these pathways, including glycolysis, the pentose phosphate pathway, TCA cycle, and the L-arabinose/D-xylose oxido-reductive pathway exhibit multiple genes for

some individual steps (Flipphi et al. FGB-08-100) . Phylogenetic analysis shows that these gene duplications were early events in the evolution of the aspergilli and it can be assumed that they have aided their proliferation in specific habitats. In addition to the pyruvate dehydrogenase and 2-oxo-glutarate dehydrogenase complexes, a third 2-oxo-acid dehydrogenase complex was discovered, and this is absent from yeast. The loci involved in this third complex are AN1726, AN3639 and AN8559. This enzyme is likely to be involved in the degradation of branched-chain aliphatic amino acids valine and isoleucine (like the orthologous complex in humans). A gene encoding glucose oxidase was found, suggesting an oxidative catabolic pathway for glucose. However, *A. nidulans* appears to lack the hydrolysing enzyme to linearise glucono-lactone. Moreover, since mutual disruption of the function of the genes for hexokinase (*hvkA*) and glucokinase (*glkA*) effectively abolishes growth on glucose (Flipphi et al., 2003), it is unlikely that glucose oxidation could lead to growth. Interestingly, *A. nidulans* appears distinct from all other aspergilli in lacking inositol-phosphate phosphatase – catalyzing the last step of inositol biosynthesis – which is non-essential in *S. cerevisiae* (Lopez et al., 1999). The encoding gene is found in synteny with its direct genetic environment in all *Aspergillus* species except *A. nidulans*. In the latter, there is no space between the neighbouring genes, AN9409 and AN9410, while in the other seven public *Aspergillus* genomes, the structurally well-conserved inositol biosynthesis gene resides between the orthologues of AN9409 and AN9410.

Two loci involved in central carbon metabolism were encountered in the *A. nidulans* genome that appear to result from horizontal gene transfer. A predicted cytosolic 2-methylcitrate dehydratase-like protein (AN1619) seems to be a prominent example (Flipphi et al. FGB-08-100). Another such event might have endowed *A. nidulans* with pyruvate-water dikinase (locus AN5843) (Flipphi and Kubicek, unpublished). In bacteria, this enzyme converts pyruvate into phosphoenolpyruvate – the reverse reaction of pyruvate kinase – thereby enabling futile cycling between the two ultimate intermediates of glycolysis. It is possible that the associated net conversion of ADP to AMP provides a physiological signal for general carbon metabolic control in *A. nidulans*.

e) Nitrogen and amino-acid metabolism

Nitrogen metabolism in *A. nidulans* has been the subject of extensive research over recent decades (Pontecorvo et al., 1953; Caddick, 2004). This has provided the model system of the GATA transcription factor AreA (AN8667), and further pathway-specific regulators and thus insight into control mechanisms for gene expression, with attention now turning to post-transcriptional control (Caddick et al., 2006). As a consequence, a large number of nitrogen metabolite transport and assimilation systems could be annotated since they had already been identified and analysed in functional detail. Similarly, the pathways of fungal aminoacid biosynthesis have been the subject of substantial previous study, including use as nutritional markers, and many could also be annotated.

Blast analysis identified signatures in several other genes which suggested putative roles in nitrogen metabolite transport, assimilation or aminoacid biosynthesis. These included a putative amino-acid permease (AN7392), purine transporter (AN7955), an additional acetamidase (AN9138) and four possible additional members of the arginase family (AN7488, AN6869, AN7669, AN3965). Bioinformatic analysis predicted functions for AN7488 and AN6869 as agmatinases and thus roles in polyamine biosynthesis. During searches for genes with amino-acid biosynthetic function, several aromatic aminotransferases were identified (AN6338 AN5041 AN8172 AN4156), at least one of which is a candidate for the final step in tyrosine biosynthesis. In addition, homology to the histidinol dehydrogenase region of the trifunctional *hisDEI* (AN0797) was identified in AN2723, but the other domains of this protein were absent. Functional analysis will be required to determine whether these genes indeed have roles in nitrogen assimilation, amino-acid metabolism or perhaps in secondary metabolism.

f) Sulphur metabolism

Almost all known genes related to eukaryotic sulphur metabolism have been annotated, and assigned an appropriate function. Within the last two years, the *astA* gene encoding an alternative sulphur transporter has been identified and characterized. The gene's

expression is strongly regulated by sulphur metabolite repression, which suggests that AstA protein is a specific sulphate transporter. The protein is distinct from known sulphate permeases and belongs to a poorly characterized large Dal5 family of allantoin transporters. Interestingly, *astA* appears to be a nonfunctional gene in some strains of *A. nidulans*. Its orthologues are present in only a few fungal species, in particular in plant pathogenic/saprophytic fungi (Pitsyk et al., 2007).

A group of genes encoding enzymes directly or indirectly implicated in homocysteine metabolism are induced by this amino acid. Coordinated regulation of this “homocysteine regulon” (Sienko et al., 2007) may prevent homocysteine reaching toxic concentrations within the cell. Recent evidence suggests that this response to homocysteine may be part of a general reaction to stress that also involves the unfolded protein response. While the molecular mechanisms involved in regulation of the “homocysteine regulon” remain to be elucidated, it is clear that the response is not part of the sulphur metabolite repression system (SMR), which controls a number of sulphur-related genes, particularly those implicated in sulphate assimilation.

g) Secondary metabolism

Secondary metabolism is the hallmark of most filamentous ascomycetes, and one of the intriguing findings arising from the genome sequencing projects is how few orthologous genes and gene clusters are shared between the different *Aspergillus* species (Nierman et al., 2005). The functions of most of the new secondary metabolic clusters revealed by the genome project remain unknown, and their biosynthetic products cannot be predicted from sequence alone. Nevertheless, the availability of the genome sequence has stimulated functional analysis studies. This has already led to the discovery of a number of metabolic products not previously reported for *A. nidulans*, namely aspyridones A and B (Bergmann et al., 2007), terrequinone A (Bok et al., 2006a) and the emericellamides (Chiang et al., 2008). An analysis of predicted non-ribosomal peptide synthetase genes has been carried out (von Döhren FGB-08-122)

h) Cytochrome P450

The extensive metabolic capacity of *A. nidulans* is also reflected in the 119 cytochrome P450 (CYP) genes (including 8 pseudogenes) that have now been annotated (Kelly et al. FGB-08-98), both manually and by exploiting the query functions of the e-Fungi data warehouse (Cornell et al., 2007). The functions of 13 of these have been determined to date, including genes for ergosterol and dityrosine biosynthesis. 32 of the genes were located close to known secondary metabolic genes such as those encoding nonribosomal peptide synthetases and polyketide synthases, probably reflecting their role in decorating secondary metabolites. Also identified were 8 putative NADPH cytochrome P450 reductase sequences. The protein encoded by AN0595 on chromosome VIII has 91% amino-acid sequence identity to the CprA protein from *A. niger*, 88% identity to the CprA from *A. fumigatus*, and is the most probable candidate reductase for the majority of the CYPs.

i) Peroxisomes

Fungal microbodies (peroxisomes, Woronin bodies) are inducible organelles that proliferate or are degraded (via an autophagy-related process termed pexophagy) in response to nutritional cues. Proteins involved in microbody biogenesis/proliferation are designated peroxins and are encoded by *PEX* genes, while genes involved in autophagy (including pexophagy) are known as *ATG* genes. A reappraisal of the *A. nidulans* genome for the presence of *PEX* and *ATG* genes has identified a number of previously missed genes. This analysis has led to the conclusion that the basic set of genes involved in microbody biogenesis, proliferation, and degradation are conserved in *A. nidulans* (Kiel and van der Klei FGB-08-78). The major differences between filamentous ascomycetes like *A. nidulans* and other organisms appears to be an aberrant RING structure on the peroxin Pex2p (part of the so-called "importomer"), the presence of a Pex16p orthologue (previously identified mainly in higher eukaryotes) and the novel protein Pex14/17p (related to the yeast-specific peroxin Pex17p), as well as multiple Pex11p paralogues (a feature seen mainly in higher eukaryotes).

With respect to pexophagy, a receptor protein required to link microbodies destined for

degradation to the autophagy machinery could not be identified in *A. nidulans*. Nevertheless, *A. nidulans* contains an Atg11p orthologue, a protein associated with selective autophagy, implying that selective microbody degradation occurs in this filamentous fungus. Because both microbody biogenesis / proliferation and autophagy / pexophagy have features that more closely resemble organelle formation / degradation in mammals rather than yeast, filamentous fungi like *A. nidulans* are ideal model systems of peroxisome homeostasis in man.

j. Cell Wall Genes

The hyphal walls of *A. nidulans* consist of an internal electron-transparent layer surrounded by an electron-dense outer layer (Jeong et al., 2004). This is consistent with the notion that the internal layer mainly contains stress-bearing polysaccharides, whereas the outer layer is enriched with glycoproteins (De Groot et al., 2005). The inventory of cell wall genes reveals the presence of genes encoding polysaccharide synthases, including genes responsible for the production of the stress-bearing polysaccharides 1,3- α -/1,4- α -glucan, 1,3- β -glucan, and chitin (De Groot et al. FGB-08-116). In addition, *A. nidulans* contains a gene (AN8444, designated as *celA*) the product of which shows similarity to bacterial plant and cellulose synthases and is predicted to be responsible for the production of 1,3- β -/1,4- β -glucan. This agrees with the presence of a linear 1,3- β -/1,4- β -glucan in the alkali-insoluble fraction of the hyphal wall of *A. fumigatus* (Fontaine et al., 2000). Importantly, four predicted GPI-protein-encoding genes were identified that are believed to have transglucosidase activity, and to act on 1,3- α -glucan (AN3308, AN4507, AN6324) or on 1,4- β -glucan (AN2385) (van der Kaaij et al., 2007; Yuan et al., 2008b). Whereas many GPI-proteins are predominantly located in the plasma membrane, others are incorporated in the fungal cell wall (reviewed in (De Groot et al., 2005)). Mass spectrometric analysis of a tryptic digest obtained by 'cell wall shaving' (Yin et al., 2008) showed the presence of twelve proteins, including ten predicted GPI-proteins. Hydrophobins are small amphipathic proteins that are only found in mycelial species and are deposited on the cell surface of conidia/ascospores, basidiocarps and aerial hyphae to render them hydrophobic (Wösten, 2001). Six putative hydrophobin-encoding genes were identified in the *A. nidulans* genome, including the already known *rodA* and *dewA*. Finally, some genes involved in spore wall maturation were identified, including some encoding chitin deacetylases, which are predicted to convert chitin into chitosan. In addition, two predicted cytosolic proteins were found that show similarity to the ScDit proteins. These synthesize precursors of the dityrosine-containing macromolecule found in the outer wall layer of the ascospores of *S. cerevisiae* (Briza et al., 1994).

k. Secretion

The secretion of proteins is important to the natural lifestyles of the aspergilli and becomes critical when those species are exploited as hosts for the commercial secretion of heterologous proteins. Their natural lifestyles make the aspergilli more effective secretors of hydrolytic enzymes than *S. cerevisiae*, suggesting that the aspergilli might have a better capacity for secretion of proteins in general. That is confirmed by the experience of the biotechnology industry, although *S. cerevisiae* can be improved to secrete some proteins to commercially acceptable yields. Therefore, the main question facing the annotation team was: does the genome annotation provide any clues as to why the aspergilli might be better protein secretors than the model fungus *S. cerevisiae* and might they be better suited to expression of proteins with human-like glycosylation? The annotation alone does not provide simple answers to these questions (Geysens et al. FGB-08-112; Tsang et al. FGB-08-121). The core facets of the chaperone cycle, formation of disulphide bonds and the unfolded protein response (UPR) show a high degree of similarity between species. Even so, differences between the aspergilli and the yeasts can be seen in the nucleotide exchange reaction within the chaperone cycle, the protein disulphide isomerase protein family and the activation mechanisms of the UPR mediated by the Hac bZip transcription factor. The functional consequences of these differences remain to be assessed.

In eukaryotes, N-glycosylation is performed via the transfer of a lipid-linked precursor structure ($\text{Glc}_3\text{Man}_9\text{GlcNAc}_2$) onto a nascent polypeptide chain within the lumen of the ER. In the aspergilli, orthologues were identified for almost all genes involved in the generation of activated sugar donors as well as the synthesis and transfer of the precursor structure to a protein backbone. Also, the genes involved in the ER-processing of protein-linked N-glycans seem to be present. However, differences were observed with *S. cerevisiae* and in some cases this makes the aspergillus system more like that of mammals, e.g. the presence of a reglucosylation enzyme (UGGT), involved in the calnexin-based glycoprotein quality control. In general, the N-glycans on proteins secreted by the aspergilli are of the high-mannose type, sometimes decorated with monosaccharides such as Galf. In contrast to *S. cerevisiae*, the aspergilli rarely synthesize hyperglycosyl structures. Nevertheless, they possess several orthologues of yeast mannosyltransferases that are known to be involved in

the process of hypermannosylation. Several orthologues for mannosidases of family 47 and 92 were identified but it remains to be evaluated whether they act upon N-glycans inside (as in the mammalian Golgi apparatus) or outside the cell. Golgi-based demannosylation would again provide the aspergillus glycosylation pathway with mammalian similarity.

I. Polar growth

The signaling pathways involved in morphogenesis and calcium responses are reasonably well-conserved when compared to those of the well-studied yeast models (Harris et al. FGB-08-135). However, *A. nidulans* possesses many additional genes implicated in morphogenesis, development, and calcium signaling that are not found in yeast. The function of most of these genes remains a mystery, though it seems likely that they will contribute to unique aspects of hyphal growth (i.e., highly polarized growth) and development (conidiation, cleistothecium formation). Notably, *A. nidulans* shares with migratory animal cells and neurons several genes involved in actin organization that are otherwise absent from yeast. The number of genes potentially implicated in hyphal morphogenesis is therefore rather large and may exceed initial expectations.

4. Data availability

All data generated from this project – refined assembly, gene structures and annotation - have been deposited within CADRE (<http://www.cadre-genomes.org.uk>). CADRE (Central *Aspergillus* Data Repository) is a public resource that provides web-based tools for visualising and analysing genomic features identified within aspergilli. These tools offer simple displays for viewing annotation assigned to predicted genes (e.g. gene symbol, public loci and GO terms) and to their protein products (e.g. family and domain similarity matches), as well as complex displays for viewing genes and other features (e.g. repeated sequences) in the context of an assembly. Using the customised search facilities provided

by CADRE, all *A. nidulans* genes can be easily sought and identified by their unique public locus identifier (AN****.4).

Acknowledgements

We acknowledge financial support by the European Commission under contract LSSG-CT-2005-018964. MC, and the use of the e-Fungi data warehouse, was supported by a grant to SGO and others as part of the BBSRC's Bioinformatics and e-Science programme II. We wish to thank Dr Michael Anderson for his input on the assembly, whilst at The University of Manchester. We also acknowledge Todd Creasy for Manatee set-up and support, currently at IGS; Brian Haas for data management and computational support (annotation), currently at the Broad; Joshua Orvis for data management and computational support (annotation), currently at IGS; Jonathan Crabtree for data management and computational support (orthologues), currently at IGS. Sandra M.J. Langeveld is acknowledged for her assistance in preparing part of the manuscript

References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., 1990. Basic local alignment search tool. *J Mol Biol.* 215, 403-410.
- Asai, K., Itou, K., Ueno, Y., Yada, T., 1998. Recognition of human genes by stochastic parsing. *Pacific Symposium on Biocomputing.* 228-39.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics.* 25, 25-29.
- Bairoch, A., 2000. The ENZYME database in 2000. *Nucleic Acids Research.* 28, 304-305.
- Bendtsen, J. D., Nielsen, H., von Heijne, G., Brunak, S., 2004. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol.* 340, 783-795.
- Bergmann, S., Schumann, J., Scherlach, K., Lange, C., Brakhage, A. A., Hertweck, C., 2007. Genomics-driven discovery of PKS-NRPS hybrid metabolites from *Aspergillus nidulans*. *Nature Chemical Biology.* 3, 213-217.
- Bernal, A., Crammer, K., Hatzigeorgiou, A., Pereira, F., 2007. Global discriminative learning for higher-accuracy computational gene prediction. *Plos Computational Biology.* 3, 488-497.
- Birney, E., Clamp, M., Durbin, R., 2004. GeneWise and Genomewise. *Genome Research.* 14, 988-995.
- Bok, J. W., Hoffmeister, D., Maggio-Hall, L. A., Murillo, R., Glasner, J. D., Keller, N. P., 2006a. Genomic mining for *Aspergillus* natural products. *Chemistry & Biology.* 13, 31-37.
- Bok, J. W., Noordermeer, D., Kale, S. P., Keller, N. P., 2006b. Secondary metabolic gene cluster silencing in *Aspergillus nidulans*. *Molecular Microbiology.* 61, 1636-1645.

- Brakhage, A. A., Sprote, P., Al-Abdallah, Q., Gehrke, A., Plattner, H., Tuncher, A., Regulation of penicillin biosynthesis in filamentous fungi. *Molecular biotechnology of fungal beta-lactam antibiotics and related peptide synthetases*, Advances in Biochemical Engineering and Biotechnology, 2004, pp. 45-90.
- Briza, P., Eckerstorfer, M., Breitenbach, M., 1994. The Sporulation-Specific Enzymes Encoded by the Dlt1 and Dlt2 Genes Catalyze a 2-Step Reaction Leading to a Soluble L-Dityrosine-Containing Precursor of the Yeast Spore Wall. *Proceedings of the National Academy of Sciences of the United States of America*. 91, 4524-4528.
- Brody, H., Griffith, J., Cuticchia, A. J., Arnold, J., Timberlake, W. E., 1991. Chromosome-Specific Recombinant-DNA Libraries from the Fungus *Aspergillus nidulans*. *Nucleic Acids Research*. 19, 3105-3109.
- Caddick, M. X., 2004. Nitrogen regulation in mycelial fungi. In: Brambl, R. Marzluf, G.A. Eds., *The Mycota III Biochemistry and Molecular Biology*, Springer-Verlag, Berlin-Heidelberg, pp. 349 - 367.
- Caddick, M. X., Jones, M. G., van Tonder, J. M., Le Cordier, H., Narendja, F., Strauss, J., Morozov, I. Y., 2006. Opposing signals differentially regulate transcript stability in *Aspergillus nidulans*. *Molecular Microbiology*. 62, 509-519.
- Cazelle, B., Pokorska, A., Hull, E., Green, P. M., Stanway, G., Scazzocchio, C., 1998. Sequence, exon-intron organization, transcription and mutational analysis of *prnA*, the gene encoding the transcriptional activator of the *prn* gene cluster in *Aspergillus nidulans*. *Molecular Microbiology*. 28, 355-370.
- Chiang, Y. M., Szewczyk, E., Nayak, T., Davidson, A. D., Sanchez, J. F., Lo, H. C., Ho, W. Y., Simityan, H., Kuo, E., Praseuth, A., Watanabe, K., Oakley, B. R., Wang, C. C. C., 2008. Molecular genetic mining of the *Aspergillus* secondary metabolome: Discovery of the emericellamide biosynthetic pathway. *Chemistry & Biology*. 15, 527-532.
- Clutterbuck, A. J., 1973. Gene symbols in *Aspergillus nidulans*. *Genetical Research Cambridge*. 21, 291-296.
- Clutterbuck, A. J., Farman, M., *Aspergillus nidulans* linkage map and genome sequence: closing gaps and adding telomeres. In: G. H. Goldman, S. A. Osmani, Eds.), *The Aspergilli: genomics, medical aspects, biotechnology, and research methods*. CRC Press, Boca Raton, 2008, pp. 57-65.
- Clutterbuck, A. J., Kapitonov, V. V., Jurka, J., Transposable elements and repeat-induced point mutation in *Aspergillus nidulans*, *Aspergillus fumigatus* and *Aspergillus oryzae*. In: G. H. Goldman, A. H. Osmani, Eds.), *The Aspergilli: genomics, medical aspects, biotechnology, and research methods* CRC Press, Boca Raton, 2008, pp. 343-355.
- Colot, H. V., Park, G., Turner, G. E., Ringelberg, C., Crew, C. M., Litvinkova, L., Weiss, R. L., Borkovich, K. A., Dunlap, J. C., 2006. A high-throughput gene knockout procedure for *Neurospora* reveals functions for multiple transcription factors. *Proceedings of the National Academy of Sciences of the United States of America*. 103, 10352-10357.
- Cornell, M. J., Alam, I., Soanes, D. M., Wong, H. M., Hedeler, C., Paton, N. W., Rattray, M., Hubbard, S. J., Talbot, N. J., Oliver, S. G., 2007. Comparative genome analysis across a kingdom of eukaryotic organisms: Specialization and diversification in the Fungi. *Genome Research*. 17, 1809-1822.
- Cultrone, A., Dominguez, Y. R., Drevet, C., Scazzocchio, C., Fernandez-Martin, R., 2007. The tightly regulated promoter of the *xanA* gene of *Aspergillus nidulans* is included in a helitron. *Molecular Microbiology*. 63, 1565-1567.
- De Groot, P. W. J., Ram, A. F., Klis, F. M., 2005. Features and functions of covalently linked proteins in fungal cell walls. *Fungal Genetics and Biology*. 42, 657-675.

- DeCaprio, D., Vinson, J. P., Pearson, M. D., Montgomery, P., Doherty, M., Galagan, J. E., 2007. Conrad: Gene prediction using conditional random fields. *Genome Research*. 17, 1389-1398.
- Fedorova, N. D., Khaldi, N., Joardar, V. S., Maiti, R., Amedeo, P., Anderson, M. J., Crabtree, J., Silva, J. C., Badger, J. H., Albarraq, A., Angiuoli, S., Bussey, H., Bowyer, P., Cotty, P. J., Dyer, P. S., Egan, A., Galens, K., Fraser-Liggett, C. M., Haas, B. J., Inman, J. M., Kent, R., Lemieux, S., Malavazi, I., Orvis, J., Roemer, T., Ronning, C. M., Sundaram, J. P., Sutton, G., Turner, G., Venter, J. C., White, O. R., Whitty, B. R., Youngman, P., Wolfe, K. H., Goldman, G. H., Wortman, J. R., Jiang, B., Denning, D. W., Nierman, W. C., 2008. Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. *PLoS Genet*. 4, e1000046.
- Felenbok, B., Flipphi, M., Nikolaev, I., 2001. Ethanol catabolism in *Aspergillus nidulans*: a model system for studying gene regulation. *Prog. Nucleic Acid Res. Mol. Biol.* 69, 149-204.
- Finn, R. D., Tate, J., Mistry, J., Coghill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L., Bateman, A., 2008. The Pfam protein families database. *Nucleic Acids Research*. 36, D281-288.
- Flipphi, M., van de Vondervoort, P.J.I., Ruijter, G.J.G., Visser, J., Arst, H.N., Jr., Felenbok, B., 2003. Onset of carbon catabolite repression in *Aspergillus nidulans*. Parallel involvement of hexokinase and glucokinase in sugar signaling. *Journal of Biological Chemistry*. 278, 11849-11857.
- Flipphi, M., Robellet, X., Dequier, E., Leschelle, X., Felenbok, B., Vélot, C., 2006. Functional analysis of *alcS*, a gene of the *alc* cluster in *Aspergillus nidulans*. *Fungal Genetics and Biology* 43, 247-260.
- Fontaine, T., Simenel, C., Dubreucq, G., Adam, O., Delepierre, M., Lemoine, J., Vorgias, C. E., Diaquin, M., Latge, J. P., 2000. Molecular organization of the alkali-insoluble fraction of *Aspergillus fumigatus* cell wall. *Journal of Biological Chemistry*. 275, 27594-27607.
- Galagan, J. E., Calvo, S. E., Cuomo, C., Ma, L. J., Wortman, J. R., Batzoglou, S., Lee, S. I., Basturkmen, M., Spevak, C. C., Clutterbuck, J., Kapitonov, V., Jurka, J., Scazzocchio, C., Farman, M., Butler, J., Purcell, S., Harris, S., Braus, G. H., Draht, O., Busch, S., D'Enfert, C., Bouchier, C., Goldman, G. H., Bell-Pedersen, D., Griffiths-Jones, S., Doonan, J. H., Yu, J., Vienken, K., Pain, A., Freitag, M., Selker, E. U., Archer, D. B., Penalva, M. A., Oakley, B. R., Momany, M., Tanaka, T., Kumagai, T., Asai, K., Machida, M., Nierman, W. C., Denning, D. W., Caddick, M., Hynes, M., Paoletti, M., Fischer, R., Miller, B., Dyer, P., Sachs, M. S., Osmani, S. A., Birren, B. W., 2005b. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature*. 438, 1105-1115.
- Ganley, A. R. D., Kobayashi, T., 2007. Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Research*. 17, 184-191.
- Gomi, K., Akeno, T., Minetoki, T., Ozeki, K., Kumagai, C., Okazaki, N., Iimura, Y., 2000. Molecular cloning and characterization of a transcriptional activator gene, *amyR*, involved in the amyolytic gene expression in *Aspergillus oryzae*. *Bioscience Biotechnology and Biochemistry*. 64, 816-827.
- Gotoh, O., 2000. Homology-based gene structure prediction: simplified matching algorithm using a translated codon (tron) and improved accuracy by allowing for long gaps. *Bioinformatics*. 16, 190-202.

- Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., Salzberg, S. L., White, O., 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*. 31, 5654-5666.
- Hull, E. P., Green, P. M., Arst, H. N., Scazzocchio, C., 1989. Cloning and Physical Characterization of the L-Proline Catabolism Gene Cluster of *Aspergillus nidulans*. *Molecular Microbiology*. 3, 553-559.
- Jeong, H. Y., Chae, K. S., Whang, S. S., 2004. Presence of a mannoprotein, MnpAp, in the hyphal cell wall of *Aspergillus nidulans*. *Mycologia*. 96, 52-56.
- Jinks, J. L., Caten, C. E., Simchen, G., Croft, J. H., 1966. Heterokaryon incompatibility and variation in wild populations of *Aspergillus nidulans*. *Heredity*. 21, 227-239.
- Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S. L., 2004. Versatile and open software for comparing large genomes. *Genome Biology*. 5, R12.
- Lamb, H. K., Hawkins, A. R., Smith, M., Harvey, I. J., Brown, J., Turner, G., Roberts, C. F., 1990. Spatial and Biological Characterization of the Complete Quinic Acid Utilization Gene-Cluster in *Aspergillus nidulans*. *Molecular & General Genetics*. 223, 17-23.
- Li, W., Rehmeyer, C. J., Staben, C., Farman, M. L., 2005. TERMINUS--Telomeric End-Read Mining IN Unassembled Sequences. *Bioinformatics*. 21, 1695-1698.
- Lopez, F., Leube, M., Gil-Mascarell, R., Navarro-Aviñó, J.P., Serrano, R., 1999. The yeast inositol monophosphatase is a lithium- and sodium-sensitive enzyme encoded by a non-essential gene pair. *Molecular Microbiology*. 31, 1255-1264.
- Mabey, J. E., Anderson, M. J., Giles, P. F., Miller, C. J., Attwood, T. K., Paton, N. W., Bornberg-Bauer, E., Robson, G. D., Oliver, S. G., Denning, D. W., 2004. CADRE: the Central *Aspergillus* Data REpository. *Nucleic Acids Research*. 32, D401-405.
- MacPherson, S., Larochele, M., Turcotte, B., 2006. A fungal family of transcriptional regulators: the zinc cluster proteins. *Microbiology and Molecular Biology Reviews*. 70, 583-604
- Machida, M., Asai, K., Sano, M., Tanaka, T., Kumagai, T., Terai, G., Kusumoto, K., Arima, T., Akita, O., Kashiwagi, Y., Abe, K., Gomi, K., Horiuchi, H., Kitamoto, K., Kobayashi, T., Takeuchi, M., Denning, D. W., Galagan, J. E., Nierman, W. C., Yu, J., Archer, D. B., Bennett, J. W., Bhatnagar, D., Cleveland, T. E., Fedorova, N. D., Gotoh, O., Horikawa, H., Hosoyama, A., Ichinomiya, M., Igarashi, R., Iwashita, K., Juvvadi, P. R., Kato, M., Kato, Y., Kin, T., Kokubun, A., Maeda, H., Maeyama, N., Maruyama, J., Nagasaki, H., Nakajima, T., Oda, K., Okada, K., Paulsen, I., Sakamoto, K., Sawano, T., Takahashi, M., Takase, K., Terabayashi, Y., Wortman, J. R., Yamada, O., Yamagata, Y., Anazawa, H., Hata, Y., Koide, Y., Komori, T., Koyama, Y., Minetoki, T., Suharnan, S., Tanaka, A., Isono, K., Kuhara, S., Ogasawara, N., Kikuchi, H., 2005. Genome sequencing and analysis of *Aspergillus oryzae*. *Nature*. 438, 1157-1161.
- Majoros, W. H., Pertea, M., Antonescu, C., Salzberg, S. L., 2003. GlimmerM, Exonomy and Unveil: three ab initio eukaryotic genefinders. *Nucleic Acids Research*. 31, 3601-3604.
- Martinelli, S. D., Gene symbols. In: S. D. Martinelli, J. R. Kinghorn, Eds., *Aspergillus: 50 years on*. Elsevier, Amsterdam, 1994, pp. 825-827.
- Martinelli, S. D., Kinghorn, J. R. Eds., 1994. *Aspergillus: 50 years on*. Elsevier, Amsterdam.
- Mulder, N., Apweiler, R., 2007. InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods in Molecular Biology*. 396, 59-70.
- Nierman, W. C., Pain, A., Anderson, M. J., Wortman, J. R., Kim, H. S., Arroyo, J., Berriman, M., Abe, K., Archer, D. B., Bermejo, C., Bennett, J., Bowyer, P., Chen, D., Collins, M.,

- Coulsen, R., Davies, R., Dyer, P. S., Farman, M., Fedorova, N., Fedorova, N., Feldblyum, T. V., Fischer, R., Fosker, N., Fraser, A., Garcia, J. L., Garcia, M. J., Goble, A., Goldman, G. H., Gomi, K., Griffith-Jones, S., Gwilliam, R., Haas, B., Haas, H., Harris, D., Horiuchi, H., Huang, J. Q., Humphray, S., Jimenez, J., Keller, N., Khouri, H., Kitamoto, K., Kobayashi, T., Konzack, S., Kulkarni, R., Kumagai, T., Lafon, A., Latge, J. P., Li, W. X., Lord, A., Lu, C., Majoros, W. H., May, G. S., Miller, B. L., Mohamoud, Y., Molina, M., Monod, M., Mouyna, I., Mulligan, S., Murphy, L., O'Neil, S., Paulsen, I., Penalva, M. A., Perteau, M., Price, C., Pritchard, B. L., Quail, M. A., Rabbinowitsch, E., Rawlins, N., Rajandream, M. A., Reichard, U., Renauld, H., Robson, G. D., de Cordoba, S. R., Rodriguez-Pena, J. M., Ronning, C. M., Rutter, S., Salzberg, S. L., Sanchez, M., Sanchez-Ferrero, J. C., Saunders, D., Seeger, K., Squares, R., Squares, S., Takeuchi, M., Tekaiia, F., Turner, G., de Aldana, C. R. V., Weidman, J., White, O., Woodward, J., Yu, J. H., Fraser, C., Galagan, J. E., Asai, K., Machida, M., Hall, N., Barrell, B., Denning, D. W., 2005b. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature*. 438, 1151-1156.
- Oakley, C. E., Oakley, B. R., 1989. Identification of gamma-tubulin, a new member of the tubulin superfamily encoded by *mipA* gene of *Aspergillus nidulans*. *Nature*. 338, 662-664.
- Osmani, A. H., Oakley, B. R., Osmani, S. A., 2006. Identification and analysis of essential *Aspergillus nidulans* genes using the heterokaryon rescue technique. *Nature Protocols*. 1, 2517-2526.
- Pel, H. J., de Winde, J. H., Archer, D. B., Dyer, P. S., Hofmann, G., Schaap, P. J., Turner, G., de Vries, R. P., Albang, R., Albermann, K., Andersen, M. R., Bendtsen, J. D., Benen, J. A., van den Berg, M., Breestraat, S., Caddick, M. X., Contreras, R., Cornell, M., Coutinho, P. M., Danchin, E. G., Debets, A. J., Dekker, P., van Dijck, P. W., van Dijk, A., Dijkhuizen, L., Driessen, A. J., d'Enfert, C., Geysens, S., Goosen, C., Groot, G. S., de Groot, P. W., Guillemette, T., Henrissat, B., Herweijer, M., van den Hombergh, J. P., van den Hondel, C. A., van der Heijden, R. T., van der Kaaij, R. M., Klis, F. M., Kools, H. J., Kubicek, C. P., van Kuyk, P. A., Lauber, J., Lu, X., van der Maarel, M. J., Meulenbergh, R., Menke, H., Mortimer, M. A., Nielsen, J., Oliver, S. G., Olsthoorn, M., Pal, K., van Peij, N. N., Ram, A. F., Rinas, U., Roubos, J. A., Sagt, C. M., Schmoll, M., Sun, J., Ussery, D., Varga, J., Vervecken, W., van de Vondervoort, P. J., Wedler, H., Wosten, H. A., Zeng, A. P., van Ooyen, A. J., Visser, J., Stam, H., 2007. Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. *Nature Biotechnology*. 25, 221-231.
- Perteau, M., Lin, X., Salzberg, S. L., 2001. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Research*. 29, 1185-1190.
- Pitsyk, S., Natorff, R., Sienko, M., Paszewski, A., 2007. Sulfate transport in *Aspergillus nidulans*: A novel gene encoding alternative sulfate transporter. *Fungal Genetics and Biology*. 44, 715-725.
- Pontecorvo, G., Roper, J. A., Hemmons, L. M., Macdonald, K. D., Buffon, A. W. J., 1953. The Genetics of *Aspergillus nidulans*. *Advances in Genetics Incorporating Molecular Genetic Medicine*. 5, 141-238.
- Salamov, A. A., Solovyev, V. V., 2000. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Research*. 10, 516-522.
- Sienko, M., Natorff, R., Zielinski, Z., Hejduk, A., Paszewski, A., 2007. Two *Aspergillus nidulans* genes encoding methylenetetrahydrofolate reductases are up-regulated by homocysteine. *Fungal Genetics and Biology*. 44, 691-700.

- Sonnhammer, E. L., von Heijne, G., Krogh, A., 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol.* 6, 175-182.
- Szewczyk, E., Nayak, T., Oakley, C. E., Edgerton, H., Xiong, Y., Taheri-Talesh, N., Osmani, S. A., Oakley, B. R., 2006. Fusion PCR and gene targeting in *Aspergillus nidulans*. *Nature Protocols.* 1, 3111-3120.
- Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., Natale, D. A., 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics.* 4, 41.
- Tatusova, T. A., Madden, T. L., 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett.* 174, 247-250.
- Todd, R. B., Davis, M. A., Hynes, M. J., 2007a. Genetic manipulation of *Aspergillus nidulans*: heterokaryons and diploids for dominance, complementation and haploidization analyses. *Nature Protocols.* 2, 822-830.
- Todd, R. B., Davis, M. A., Hynes, M. J., 2007b. Genetic manipulation of *Aspergillus nidulans*: meiotic progeny for genetic analysis and strain construction. *Nature Protocols.* 2, 811-821.
- Tsuji, G., Kenmochi, Y., Takano, Y., Sweigard, J., Farrall, L., Furusawa, I., Horino, O., Kubo, Y., 2000. Novel fungal transcriptional activators, Cmr1p of *Colletotrichum lagenarium* and Pig1p of *Magnaporthe grisea*, contain Cys2His2 zinc finger and Zn(II)2Cys6 binuclear cluster DNA-binding motifs and regulate transcription of melanin biosynthesis genes in a developmentally specific manner. *Molecular Microbiology.* 38, 940-954.
- Unkles, S. E., Campbell, E. I., Punt, P. J., Hawker, K. L., Contreras, R., Hawkins, A. R., Vandenhondel, C., Kinghorn, J. R., 1992. The *Aspergillus niger* *niaD* Gene Encoding Nitrate Reductase - Upstream Nucleotide and Amino-Acid-Sequence Comparisons. *Gene.* 111, 149-155.
- van der Kaaij, R. A., Yuan, X. L., Franken, A., Ram, A. F. J., Punt, P. J., van der Maarel, A., Dijkhuizen, L., 2007. Two novel, putatively cell wall-associated and glycosylphosphatidylinositol-anchored alpha-glucanotransferase enzymes of *Aspergillus niger*. *Eukaryotic Cell.* 6, 1178-1188.
- Woloshuk, C. P., Foutz, K. R., Brewer, J. F., Bhatnagar, D., Cleveland, T. E., Payne, G. A., 1994. Molecular Characterization of Aflr, a Regulatory Locus for Aflatoxin Biosynthesis. *Applied and Environmental Microbiology.* 60, 2408-2414.
- Wortman, J. R., Fedorova, N., Crabtree, J., Joardar, V., Maiti, R., Haas, B. J., Amedeo, P., Lee, E., Angiuoli, S. V., Jiang, B., Anderson, M. J., Denning, D. W., White, O. R., Nierman, W. C., 2006a. Whole genome comparison of the *A-fumigatus* family. *Medical Mycology.* 44, S3-S7.
- Wösten, H. A. B., 2001. Hydrophobins: Multipurpose proteins. *Annual Review of Microbiology.* 55, 625-646.
- Xiang, X., Osmani, A. H., Osmani, S. A., Xin, M., Morris, N. R., 1995. *nudF*, a nuclear migration gene in *Aspergillus nidulans*, is similar to the human LIS-1 gene required for neuronal migration. *Molecular Biology of the Cell.* 6, 297-310.
- Yin, Q. Y., de Groot, P. W. J., de Koster, C. G., Klis, F. M., 2008. Mass spectrometry-based proteomics of fungal wall glycoproteins. *Trends in Microbiology.* 16, 20-26.

- Yuan, X. L., Roubos, J. A., van den Hondel, C., Ram, A. F. J., 2008a. Identification of InuR, a new Zn(II)₂Cys₆ transcriptional activator involved in the regulation of inulinolytic genes in *Aspergillus niger*. *Molecular Genetics and Genomics*. 279, 11-26.
- Yuan, X. L., van der Kaaij, R. M., van den Hondel, C., Punt, P. J., van der Maarel, M., Dijkhuizen, L., Ram, A. F. J., 2008b. *Aspergillus niger* genome-wide analysis reveals a large number of novel alpha-glucan acting enzymes with unexpected expression profiles. *Molecular Genetics and Genomics*. 279, 545-561.

ACCEPTED MANUSCRIPT

Aspergillus Species	Sequencers	Annotation methods	References	Links and Accession Numbers to Sequences
<i>A. nidulans</i> (<i>Emericella nidulans</i> FGSC A4)	Broad Institute	Calhoun annotation system, including FGENESH (Salamov, and Solovyev, 2000); FGENESH+; and GENEWISE (Birney et al., 2004)	Galagan et al. (2005)	Genbank: AACD000000000 http://www.broad.mit.edu/annotation/genome/aspergillus_group/
<i>A. fumigatus</i> (Af293)	The Institute for Genome Research (TIGR)	TIGR annotation pipeline, including GlimmerM; Exonomy; Unveil (Majoros <i>et al.</i> , 2003); and GeneSplicer (Perteau et al., 2001) plus manual annotation	Nierman et al. (2005)	Genbank: AAHF000000000 http://www.tigr.org/tdb/e2k1/afu1/

<i>A. oryzae</i> (RIB40)	National Institute of Advanced Industrial Science and Technology (AIST)/ National Institute of Technology and Evaluation (NITE)	AIST annotation pipeline including ALN (Gotoh, 2000); GeneDecoder (Asai et al., 1998); and GlimmerM	Machida et al. (2005)	DDBJ: AP007150 - AP007177 http://www.bio.nite.go.jp/dogan/MicroTop?GENOME_ID=ao
<i>A. niger</i> (CBS 513.88)	DSM Food Specialties/ Biomax Informatics AG	Pedant-Pro™ Sequence Analysis Suite plus manual annotation by <i>Aspergillus</i> research community	Pel et al. (2007)	EMBL: AM270980 - AM270998 http://www.dsm.com/en_US/html/dfs/genomics_aniger.htm/
<i>A. niger</i> (ATCC 1015)	Joint Genome Institute (JGI)	JGI annotation pipeline including FGENESH; FGENESH+; and GENEWISE		http://genome.jgi-psf.org/Aspni1/

<p><i>A. fumigatus</i> (A1163)</p>	<p>J. Craig Venter Institute (JCVI)/ Merck & Co.</p>	<p>JCVI annotation pipeline including GlimmerHMM (Majoros et al., 2004); Genezilla (Majoros et al., 2004); SNAP (Korf, 2004); Genewise; Twinscan (Flicek et al., 2003); PASA (Haas et al., 2003); and EvidenceModeler (Haas et al., 2008)</p>	<p>Fedorova et al. (2008)</p>	<p>GenBank: ABDB000000000</p>
<p><i>A. clavatus</i> (NRRL1)</p>	<p>J. Craig Venter Institute (JCVI)</p>	<p>JCVI annotation pipeline including GlimmerHMM; Genezilla; SNAP; Genewise; Twinscan; PASA; and EvidenceModeler</p>	<p>Fedorova et al. (2008)</p>	<p>GenBank: AAKD000000000 http://www.tigr.org/tdb/e2k1/acla1/</p>

<i>A. fischerianus</i> (<i>Neosartorya fischeri</i> NRRL181)	J. Craig Venter Institute (JCVI)	JCVI annotation pipeline including GlimmerHMM; Genezilla; SNAP; Genewise; Twinscan; PASA; and EvidenceModeler	Fedorova et al. (2008)	GenBank: AAKE000000000 http://www.tigr.org/tdb/e2k1/nfa1/
<i>A. terreus</i> (NIH 2624)	Broad Institute	Calhoun annotation system, including FGENESH; GENEID (Parra et al., 2000); and GENEWISE		Genbank: AAJN000000000 http://www.broad.mit.edu/annotation/genome/aspergillus_group/
<i>A. flavus</i> (NRRL 3357)	North Carolina State University/ J. Craig Venter Institute (JCVI)	JCVI annotation pipeline including GlimmerHMM; Genezilla; SNAP; Genewise; Twinscan; PASA; and EvidenceModeler		Genbank: AAIH000000000 http://www.aspergillusflavus.org/genomics/

Table 1: Genome sequences from species in the *Aspergillus* clade All publicly available *Aspergillus* genome sequences are accessible at <http://www.cadre-genomes.org.uk/>

Table 2. Overview of the twelve annotated classes of Transcription Factors (TF) in the *A. nidulans* genome.

TF class	Table in this manuscript	DNA Binding-Domain	pfam Domain	number of genes
1	Supplementary Table 1	Zn2Cys6	00172	330
2	Supplementary Table 2	C2H2+ Zn2Cys6	00096 & 00172	12
3	Supplementary Table 3	C2H2	00096	60
4	Supplementary Table 4	GATA	00320	6
5	Supplementary Table 5	b-ZIP	07716	22
6	Supplementary Table 6	b-HLH	00010	12
7	Supplementary Table 7	CBF/CCAAT	00800	6
8	Supplementary Table 8	APSES	02292	4
9	Supplementary Table 9	Myb-like	00249	12
10	Supplementary Table 10	Fork head domain	pfam00250 and 00498	4
11	Supplementary Table 11	Homeodomain-box	00046	5
12	Supplementary Table 12	MADS-box	00319	2
	Supplementary Table 13	Miscellaneous Zn2Cys6		15