Medical Mycology September 2006, **44**, S3–S7

Taylor & Francis
Taylor & Francis Group

Whole genome comparison of the *A. fumigatus* family

J. R. WORTMAN*, N. FEDOROVA*, J. CRABTREE*, V. JOARDAR*, R. MAITI*, B. J. HAAS*, P. AMEDEO*, E. LEE*, S. V. ANGIUOLI*, B. JIANG†, M. J. ANDERSON‡, D. W. DENNING‡, O. R. WHITE* & W. C. NIERMAN*

**The Institute for Genomic Research, Rockville, MD, USA, †Merck & Co., Inc., Rahway, NJ, USA, and ‡University of Manchester, Manchester, UK*

The availability of the genome sequences of multiple *Aspergillus* spp. presents the research community with an unprecedented opportunity for discovery. The genomes of *Neosartorya fischeri* and *Aspergillus clavatus* have been sequenced in order to extend our knowledge of *Aspergillus fumigatus*, the primary cause of invasive aspergillosis. Through comparative genomic analysis, we hope to elucidate both obvious and subtle differences between genomes, developing new hypotheses that can be tested in the laboratory. A preliminary examination of the genomes and their predicted proteomes reveals extensive conservation between protein sequences and significant synteny, or conserved gene order. Comparative genomic analysis at the level of these closely related aspergilli should provide important insight into the evolutionary forces at play and their effect on gene content, regulation and expression.

Keywords *Aspergillus*, aspergillosis, genome, synteny

Introduction

The publication of the genomes of *Aspergillus fumigatus*, a human pathogen, *Aspergillus oryzae*, important in food production, and *Aspergillus nidulans*, a genetic model organism, represents a milestone for the *Aspergillus* research community, expanding knowledge of fungal physiology and mechanisms of gene regulation [1–3]. Although members of the same genus, phylogenetic analysis of these species revealed that the three aspergilli differ considerably in their genome sequences. Predicted orthologs shared by all three species display an average of only 68% amino acid identity, comparable to that between mammals and fish. The three species also differ considerably in genome size and show extensive structural reorganization [3].

Given the large evolutionary distance between the sequenced aspergilli, three additional genome projects are being funded by the National Institute of Allergy and Infectious Disease, National Institutes of Health (NIAID, USA) with the goal of better elucidating the genome of pathogenic *A. fumigatus*: *Neosartorya fischeri* (*Aspergillus fischerianus*), *Aspergillus clavatus* and *Aspergillus terreus*. The objectives in sequencing these three genomes are to use comparative genomics to improve annotation in the sequenced *Aspergillus* genomes, to provide new targets for experimental studies, to identify differences in gene content and/or regulatory elements that might contribute to pathogenicity, and to facilitate vaccine component selection with the ultimate goal of preventing invasive aspergillosis. The genomes of *N. fischeri* and *A. clavatus* are being sequenced by The Institute for Genomic Research (TIGR) and *A. terreus* is being sequenced by the Broad Institute.

Previous phylogenetic studies have identified *N. fischeri* (the teleomorph of *A. fischerianus*) as the most closely related species to *A. fumigatus* apart from the occoltum clade, *A. fumigatus var ellipticus* [4–6]. *Neosartorya fischeri*, while ubiquitously present in nature, is best known as a food-borne fungus. Its thermoresistant ascospores allow it to survive heat processing and cause spoilage of processed fruits and juices [7]. *Neosartorya fischeri* has only rarely been identified as a pathogen, although there is recent data suggesting that it may be more prevalent than believed due to misidentification in the laboratory [8].

Correspondence: J. R. Wortman, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD, USA. Tel: +301 795 7590; E-mail: jwortman@tigr.org

DOI: 10.1080/13693780600835799

Aspergillus clavatus is also a very rare human pathogen, with only one medical case, of post-surgery endocarditis, reported [9]. It has been suggested that its lack of virulence may be due to its slower growth rate at 37 °C than *A. fumigatus* and its larger spore size, which may prevent lung penetration [10]. Although not a common pathogen, it is an allergen and has been shown to be the cause of an extrinsic allergic alveolitis known as malt worker's lung [11]. *Aspergillus clavatus* produces a number of mycotoxins including patulin, kojic acid, cytochalasins and tremorgenic mycotoxins and causes neurotoxicosis in sheep and cattle fed infected grain [12,13].

The assembled genomic sequences of *N. fischeri* and *A. clavatus* are now available to the public and a first round of automated annotation has allowed the investigation of the genome structure and gene content of these species in comparison to the published *A. fumigatus* Af293. The genome sequence of a second *A. fumigatus* strain (CEA10) was made available by Merck & Co (Rahway, NJ, USA) and is also included in this preliminary work.

Comparative analysis of closely related *Aspergilli*

The genomes of *N. fischeri* NRRL 181 and *A. clavatus* NRRL 1 were sequenced by the whole genome random shotgun method, assembled using Celera assembler, and run through an automated annotation pipeline as described for *A. fumigatus* [1]. The assembled genome sizes are similar, as seen in Table 1, with *N. fischeri* approximately 10% larger. Gene numbers are approximate, as they reflect the raw output of the automated pipeline without manual review and modification, with the exception of *A. fumigatus* which was manually improved prior to publication. Due to known issues with missing and merged gene predictions, we expect that the gene number for *A. clavatus*, in particular, will increase.

Table 1 Genome statistics

	AFU Af293	AFU CEA10	NFA	ACLA
Size (Mb)	28.8	29.2	32.0	27.7
Supercontigs >2kb	19	32	256	28
GC content	49.8	49.4	49.6	49.2
Genes	9854	10 099	10 929	8653
Mean gene length (bp)	1461.8	1425.2	1434.3	1436.8
With introns (%)	78.4	78.4	78.2	83
Mean exons per gene	2.9	2.8	2.9	3.0
Genes in ortholog clusters	8769	8874	9021	7672
Proteome (%)	89	88	83	89

The predicted proteomes of the four analyzed genomes, *A. fumigatus* Af293, *A. fumigatus* CEA10, *N. fischeri* and *A. clavatus* were clustered into ortholog groups using a mutual best hit algorithm based on all vs. all BLAST results [SV Angiuoli, personal communication]. Using this method, we identified 9845 ortholog clusters, of which 7076 represent all analyzed genomes (6862 containing a single member from each proteome). The average protein identity between *A. fumigatus* and *N. fischeri* orthologs is 94%, while that between *A. fumigatus* and *A. clavatus* is 80%, supporting the published phylogeny of these genomes [14]. As previously described, the average protein identity between *A. fumigatus* and *A. oryzae* is 70%, and that between *A. fumigatus* and *A. nidulans* is 66% [3]. A set of 1000 orthologs between these five species was used to generate an inclusive phylogeny (Fig. 1).

Previously identified genes associated with pathogenicity [1, Table S1] are well conserved between *A. fumigatus*, *N. fischeri* and *A. clavatus*, suggesting that most have a critical role in cell growth and maintenance. Putative species-specific genes include those that encode extracellular proteins, transporters, transcription factors, secondary metabolic enzymes and proteins with no identifiable function.

Despite the apparent asexuality of *A. fumigatus* and *A. clavatus*, their genomes possess a full set of sexual

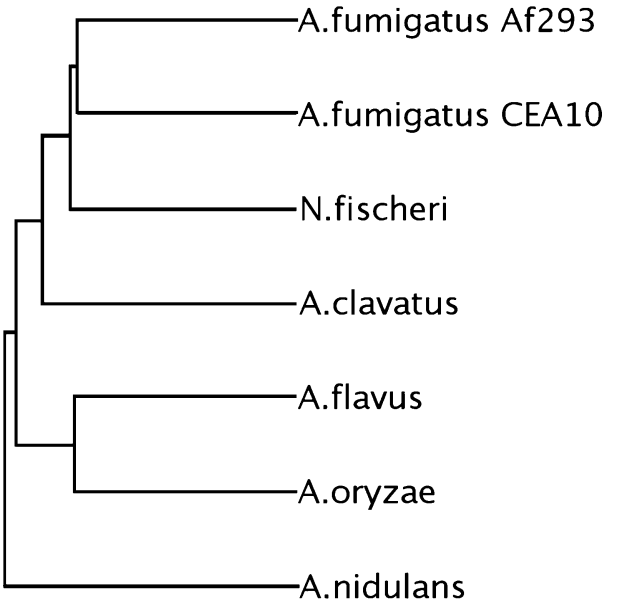

Fig. 1 Phylogenetic tree of five *Aspergillus* spp. A set of 1000 ortholog clusters containing a single representative from each of the genomes analyzed with consistent protein lengths was used to generate a phylogenetic tree. The blastP bit scores were converted to a distance metric, and the Phylip package was used to create the tree using the UPGMA method.

development genes, which is consistent with a recent loss of sexuality or a hidden sexual stage. As has been shown previously, *A. fumigatus* isolates contain either an HMG or an alpha box mating-type (MAT1-1 or MAT1-2), but never intact copies of both, suggesting that it is a heterothallic species [1,15].

Comparative analysis of the mating loci shows that *A. fumigatus* Af293 and CEA10 belong to the opposing mating types. Their mating type genes are not true idiomorphs, as they occupy adjacent positions on the chromosome. CEA10 encodes an alpha box mating-type (MAT1) protein, while the adjacent locus contains a truncated copy of the HMG-box gene. This same configuration is seen in the strain of *A. clavatus* analyzed. In contrast, the Af293 mating locus encodes an HMG box mating-type protein (MAT2), adjacent to a hypothetical protein that is distinct from the MAT1 protein. *Neosartorya fischeri*, which is homothallic, contains two mating locus regions, one of each mating type described. Notably, the locus containing the intact HMG-box gene contains a truncated DNA lyase and is surrounded by several types of transposable elements, suggesting a possible translocation event. A similar pattern was observed in the *A. nidulans* genome where the second mating locus also encoding the MAT1 mating-type protein was located in a non-syntenic region [3].

Conserved synteny and genome structure

Ortholog groups were used to determine syntenic regions between genomes, by correlating ortholog matches with their genomic locations and relative gene order. As expected, *A. fumigatus* shares larger syntenic regions with *N. fischeri* and *A. clavatus*, covering a significantly larger percentage of each genome, than with the more evolutionarily distant *A. oryzae* and *A. nidulans*. Looking at *A. fumigatus* chromosome 1 and its synteny with *N. fischeri* and *A. clavatus* (Fig. 2) illustrates the colinearity common between these genomes, with small insertions/deletions and inversions visible and less conservation at the subtelomeric ends. Transposon-related sequences, displayed as triangles, are commonly associated with interruptions in synteny. There is evidence of genome translocations at syntenic breakpoints, even between these closely related genomes, most notably between *A. fumigatus* chromosomes 2 and 5 in comparison with *N. fischeri*, and between *A. fumigatus* chromosomes 2 and 6 in comparison with *A. clavatus*. Examination of the ortholog cluster data integrated with the genomic context will allow us to improve the annotation of these genomes, and identify real differences in genome structure and gene content.

Even within the single species, *A. fumigatus* isolates display variation in their genome structure, with small-scale insertions/deletions and rearrangements. Careful inspection of the Af293 and CEA10 genomes revealed a number of unique, strain-specific genes (2%), located predominantly in non-syntenic subtelomeric blocks. A few strain-specific regions contain putative gene clusters involved in secondary metabolism, osmotolerance, or arsenic resistance, highlighting the potential role of the subtelomeric regions in maintaining species variability.

Annotation consistency affects analysis

Despite our best efforts at creating customized annotation pipelines for the *Aspergillus* genomes, utilizing state-of-the-art gene prediction programs and leveraging the wealth of fungal genome data available, there are problems and inconsistencies in the data sets that need to be addressed before a complete comparative analysis can be accurately accomplished (Fig. 3).

There is very little cDNA or EST data available for these genomes, so gene finding programs were trained on small data sets of manually curated gene models based on protein homology to public databases. The prevalence of introns in these species, at an average of 2 per gene, further complicates gene identification. Common annotation problems include missed or incorrect 5′ exons, and the inappropriate merging of neighboring genes into single, erroneous gene structures. An examination of the protein alignments of the 7076 four-way ortholog clusters reveals that only 5738 (82%) have an average sequence coverage of 90% or higher, suggesting that merged or truncated protein sequences are prevalent.

Identification of species-specific genes requires accurate ortholog associations, investigation of regulatory regions depends on accurate 5′ exons, and evolutionary studies depend on the concordance of gene structure annotation between orthologous genes. Therefore, correcting annotation errors in these, and other published fungal genomes, should be a priority. Inaccurate or missing gene models submitted to public databases present an enormous problem for many areas of contemporary biology including expression studies, functional genomics, phylogenetic analysis, and crystallography. These genomes would still benefit from the generation of additional ESTs and/or the validation of annotation through reverse transcription-polymerase chain reaction. There is also a need for more algorithm development to fully leverage comparative genomics

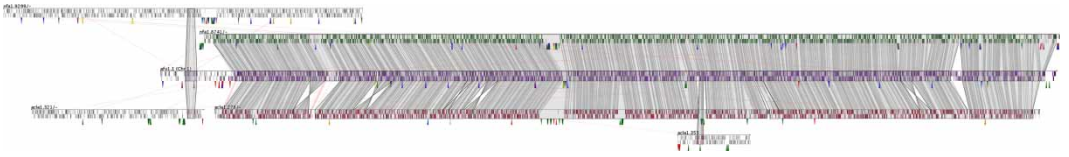

Fig. 2 *Aspergillus fumigatus* chromosome 1: Synteny with *Neosartorya fischeri* and *A. clavatus*. Ortholog clusters were computed between the *A. fumigatus* Af293, *N. fischeri* and *A. clavatus* proteomes and used to define syntenic relationships between chromosomes. Chromosome 1 of *A. fumigatus* is located in the center and is colored purple. *Neosartorya fischeri* (green) and *A. clavatus* (red) supercontigs that which share syntenic blocks with *A. fumigatus* are displayed above and below. Genes without ortholog matches in all three genomes are colored gray. The image was generated by the SYBIL comparative genomics interface [http://sybil.sourceforge.net/sybil/ (accessed on 7/11/06].

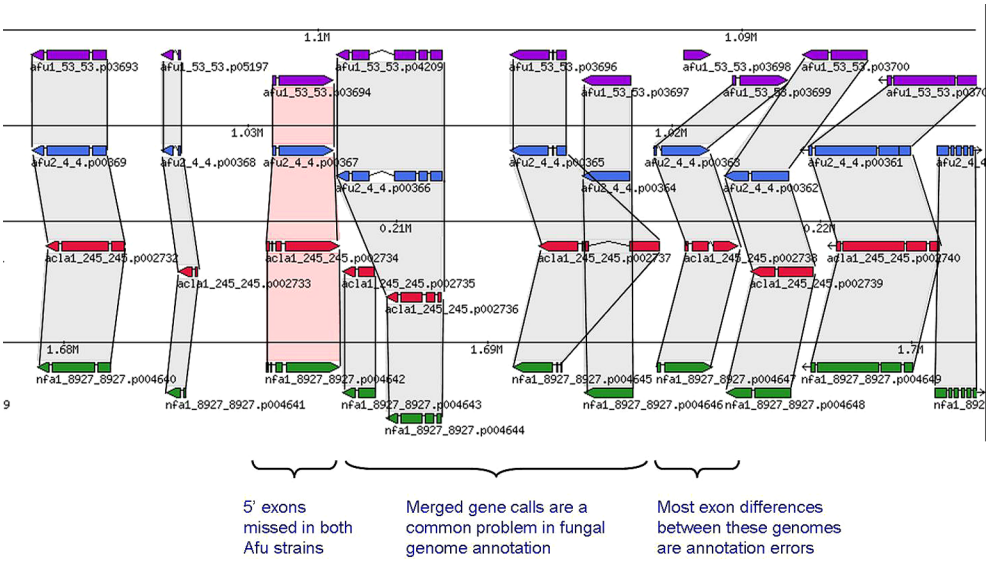

Fig. 3 Detailed synteny view highlighting annotation errors. This zoomed-in view of a syntenic block, in which individual gene structures are visible, shows some of the data inconsistencies that can adversely affect automated comparative analyses. *Aspergillus fumigatus* Af293 genes are colored purple, *A. fumigatus* CEA10 genes are colored blue, *Neosartorya fischeri* genes are colored green and *A. clavatus* genes are colored red.

data into new and improved gene models, to decrease the need for manual intervention.

Conclusion

An initial examination of the genome sequences and predicted proteomes of *A. fumigatus*, *N. fischeri* and *A. clavatus* reveals extensive conservation of gene content and order and supports the previously determined phylogeny. The core chromosome regions encode predominantly highly conserved housekeeping genes, whereas strain- and lineage-specific genes are found in subtelomeric regions. Breaks in synteny between these genomes are often flanked by repeats and transposable elements.

Further characterization of strain and species-specific genes, polymorphisms and differential regulation should foster our understanding of mechanisms of pathogenicity, environmental adaptation and resistance to anti-fungal treatments.

References

1 Nierman WC, Pain A, Anderson MJ, *et al*. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature* 2005; **438**: 1151–1156.
2 Machida M, Asai K, Sano M, *et al*. Genome sequencing and analysis of *Aspergillus oryzae*. *Nature* 2005; **438**: 1157–1161.
3 Galagan JE, Calvo SE, Cuomo C, *et al*. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* 2005; **438**: 1105–1115.
4 Pringle A, Baker DM, Platt JL, *et al*. Cryptic speciation in the cosmopolitan and clonal human pathogenic fungus *Aspergillus fumigatus*. *Evolution Int J Org Evolution* 2005; **59**: 1886–1899.
5 Varga J, Vida Z, Toth B, *et al*. Phylogenetic analysis of newly described *Neosartorya* species. *Antonie Van Leeuwenhoek* 2000; **77**: 235–239.
6 Wang L, Yokoyama K, Miyaji M, *et al*. Mitochondrial cytochrome b gene analysis of *Aspergillus fumigatus* and related species. *J Clin Microbiol* 2000; **38**: 1352–1358.

7 Beuchat LR. Extraordinary heat resistance of *Talaromyces flavus* and *Neosartorya fischeri* ascospores in fruit products. *J Food Sci* 1986; **51**: 1506–1510.
8 Balajee SA, Gribskov J, Brandt M, *et al*. Mistaken identity: *Neosartorya pseudofischeri* and its anamorph masquerading as *Aspergillus fumigatus*. *J Clin Microbiol* 2005; **43**: 5996–5999.
9 Opal SM, Reller LB, Harrington G, *et al*. *Aspergillus clavatus* endocarditis involving a normal aortic valve following coronary artery surgery. *Rev Infect Dis* 1986; **8**: 781–785.
10 Pitt JI. The current role of *Aspergillus* and *Penicillium* in human and animal health. *J Med Vet Mycol* 1994; **32**(Suppl 1): 17–32.
11 Grant IW, Blackadder ES, Greenberg M, *et al*. Extrinsic allergic alveolitis in Scottish maltworkers. *Br Med J* 1976; **1**: 490–493.
12 Varga J, Rigo K, Molnar J, *et al*. Mycotoxin production and evolutionary relationships among species of *Aspergillus* section *Clavati*. *Antonie Van Leeuwenhoek* 2003; **83**: 191–200.
13 Sabater-Vilar M, Maas RF, De Bosschere H, *et al*. Patulin produced by an *Aspergillus clavatus* isolated from feed containing malting residues associated with a lethal neurotoxicosis in cattle. *Mycopathologia* 2004; **158**: 419–426.
14 Peterson SW. Phylogenetic relationships in *Aspergillus* based on rDNA sequence analysis. In: Samson RA, Pitt JI (eds). *Integration of Modern Taxonomic Methods for Penicillium and Aspergillus Classification*. Amsterdam, The Netherlands: Harwood Academic Publishers, 2000: 323–355.
15 Paoletti M, Rydholm C, Schwier EU, *et al*. Evidence for sexuality in the opportunistic fungal pathogen *Aspergillus fumigatus*. *Curr Biol* 2005; **15**: 1242–1248.