Comparative analysis of an emerging fungal pathogen, *Aspergillus terreus*

Submitted by:

Bruce Birren The Broad Institute of Harvard and MIT – bwb@broad.mit.edu
David Denning The University of Manchester – ddenning@man.ac.uk
Bill Nierman – The Institute for Genomic Research - wnierman@tigr.org

August 26, 2004

Summary

Aspergillosis causes significant mortality and morbidity worldwide. *Aspergillus terreus*, has emerged as a significant cause of aspergillosis and its incidence is growing. Invasive aspergillosis caused by *A. terreus* carries a much higher mortality rate than any of the more than 20 pathogenic Aspergillus species, with mortality 100% in many series. Further, *A. terreus* is completely resistant to amphotericin B, a crucial treatment for fungal infections. Understanding the *A. terreus* genome and the differences between it and the genomes of carefully selected relatives will vastly improve our knowledge of pathogenesis, antibiotic resistance, and the biology of these infectious agents. Towards these ends we propose sequencing the genomes of *A. terreus*, *A. fischerianus*, and *A. clavatus*.

The objectives in sequencing these three *Aspergilli* genomes are to use comparative genomics to:

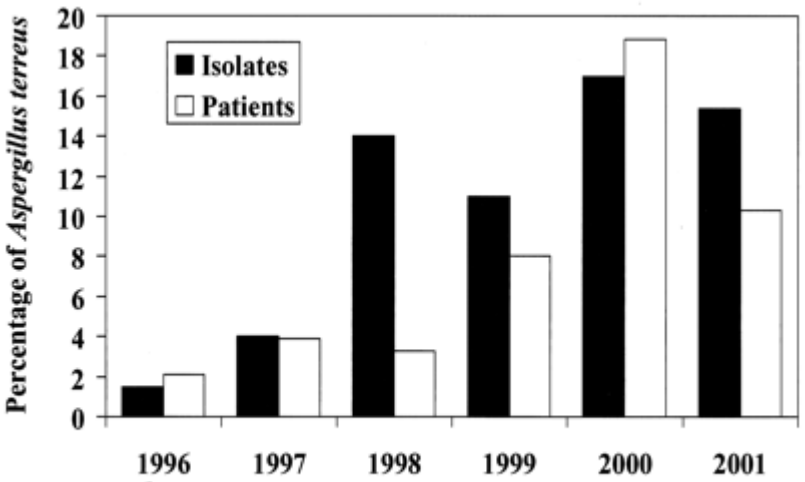
- Identify virulence determinants
- Improve annotation in *Aspergillus* genomes and provide new targets for experimental studies
- Understand amphotericin B resistance
- Facilitate vaccine component selection, with the aim of preventing invasive aspergillosis

The sequence of *A. fumigatus* has yielded a wealth of information about the genes and genome of this important class of pathogens. However, the interpretation of these data would be tremendously advanced through a comparative approach. These three genomes have been selected to support comparative studies based on their evolutionary relationship to each other as well as to exploit critical phenotypic differences. Comparative analysis of their sequence, along with that of *A. fumigatus*, will be the single most important step yet in establishing the mechanisms of infection, drug resistance and immunogenicity while propelling the development of new diagnostics, treatments and vaccines.

***Aspergillus terreus:* an emerging pathogen**

Invasive aspergillosis is the leading infectious cause of death in leukemia and stem cell transplantation, affecting thousands of patients each year in the US alone. Of the most common pathogenic species, *Aspergillus terreus* is ranked third. In a large worldwide series, it was the cause of 7% of invasive aspergillosis cases. In some centers, it is much more common (*e.g.* the MD Anderson Cancer Center (27%) and the University Hospital of Innsbruck (>30%)). An increasing frequency of infection has been documented in several sites, including the University of Alabama (Fig 1). Documented sources include hospital water and potted plants. It is certainly therefore an **'Emerging infection'**.

Fig. 1.



	1996	1997	1998	1999	2000	2001
Total *Aspergillus* isolates	66	98	134	204	179	188

Baddley *et al.*, J Clin Microbiol 2003;41:5525.

Of further concern, invasive aspergillosis caused by *A. terreus* carries a much higher mortality than any other species, with mortality 100% in many series. Moreover, it is completely resistant to amphotericin B, a first line treatment for fungal infections, and so species level diagnosis is critical for good treatment decisions. Identification can be slow, as there are no rapid tests, and the antigen tests (galactomannan and ß-glucan) do not distinguish species.

An active research community is well positioned to benefit from the *Aspergillus terreus* genome sequence. *A. terreus* has been the subject of over 500 research papers.
A. terreus has been reported to be allergenic, causing allergic bronchopulmonary aspergillosis. It is also a systemic pathogen of German shepherd dogs, causing spinal osteomyelitis, and of pregnant cows, causing abortion. *Aspergillus terreus* occupies an unusual place amongst pathogenic fungi, in also being an important industrial organism, where it is used to produce lovastatin and itaconic acid. It also produces a number of secondary metabolites and mycotoxins, including territrem A, citreoviridin, citrinin, gliotoxin, patulin, terrein, terreic acid and terretonin. The economic importance of these industrial uses and toxins further increases the number of scientists who will contribute to the analysis of the genome sequence.

A. terreus is a filamentous ascomycete and is commonly found in soil. In addition to producing typical *Aspergillus* aerial hyphae, *A. terreus* is unique among *Aspergilli* in producing lateral cells termed aleurospores in the absence of typical conidiophore structures in submerged culture (Klick and Pitt 1992). The presence of aleuorospores emphasizes the differences between the sequenced species, and could be of relevance in a clinical setting.

A. terreus has a haploid genome that is approximately 35 Mb, and has a G+C content of 50–60%. More than 30 genes have been sequenced from *A. terreus*, including 18 genes required for lovastatin biosynthesis. Most genes contain multiple, small introns. This species is highly amenable to molecular genetic manipulation with a variety of available tools, including gene libraries, transformation methods, dominant selectable markers, and gene expression systems. A clinical isolate of *A. terreus* (NIH 2624) will be sequenced. This was originally a referred isolate to the Fungus Testing Laboratory in San Antonio, Texas. It has been carefully studied in animal models and produces characteristic pathology in an invasive setting.

***Aspergillus terreus* comparative genomics**

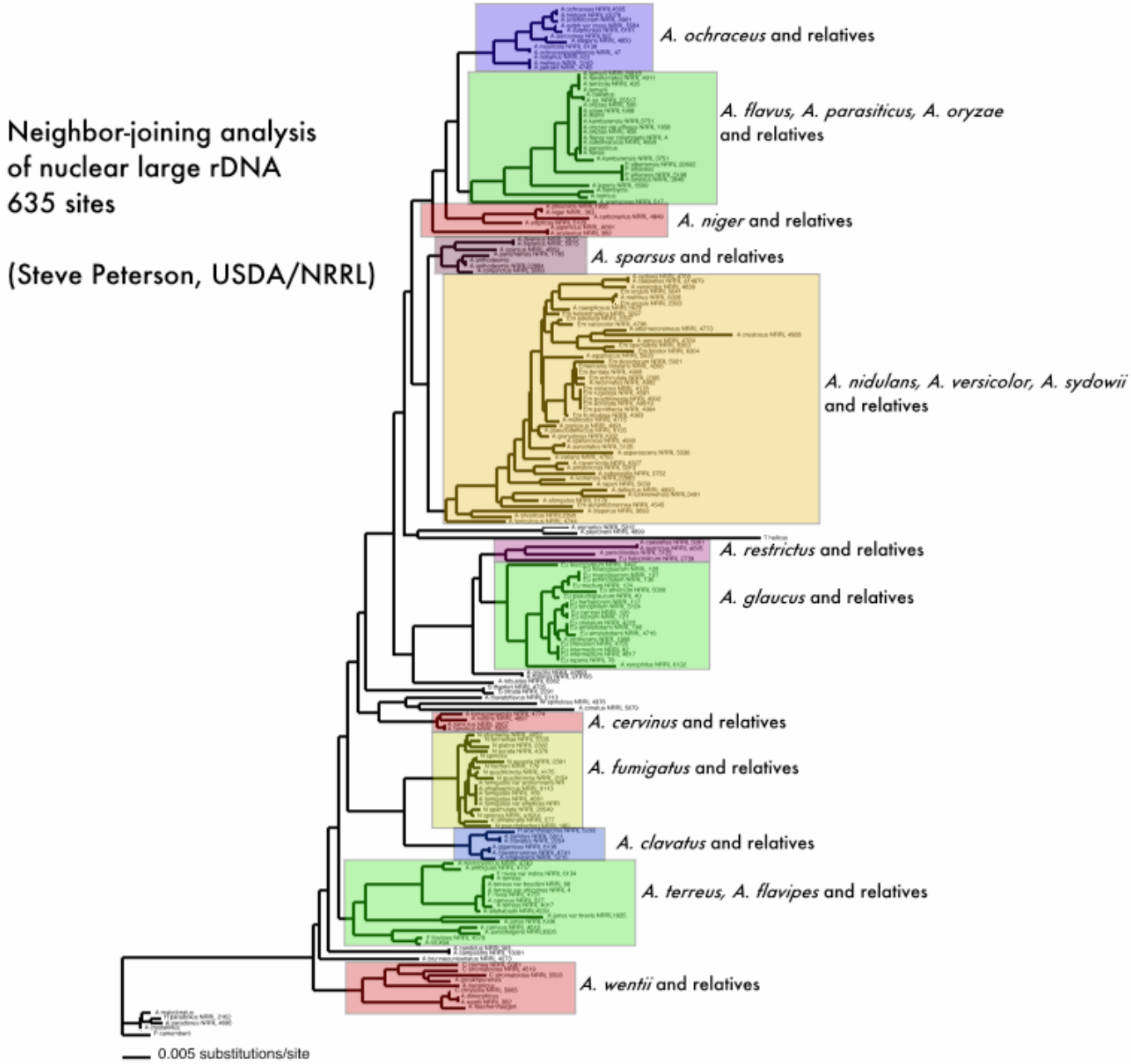
Thoughtful selection of genomes is key to the success of comparative genomics studies. Selection criteria, in turn, are dictated by the scientific goals of the project. In this case we have three primary objectives:

- To use sequence conservation among related species to improve gene annotation and recognize candidate regulatory elements in two important *Aspergillus* pathogens, *A. fumigatus* and *A. terreus*.
- To compare the genes in these different species in the context of their distinct phenotypes to dissect virulence factors in pathogenic *Aspergillus*.
- To gain further insight into sexual reproduction pathways in pathogenic *Aspergilli*, to exploit genetic tools in the lab.

Ongoing analysis has made good use of the sequence of the *Aspergillus nidulans* and *Aspergillus oryzae* genomes in comparisons with *A. fumigatus, A. oryzae* and *A. nidulans* to study gene and genome evolution among these species (Figure 3). However, these three organisms were not selected based on their value for comparative studies. In fact, and as is clear from the *Aspergillus* phylogenetic tree (Figure 2), each of these are too distant from *A. fumigatus* to effectively fulfill the above criteria. It is sobering to realize that, contrary to any assumption that organisms all named Aspergillus must be closely related, the extent of divergence between *A. fumigatus*, and *A. nidulans*, or *A. oryzae* is the same as that between humans and fish! In this light it is not surprising that only approximately 50% of each genome can be aligned with the corresponding region of each of the other two. In contrast, the positions of *A. fischerianus* and *A. clavatus* make them extremely useful for comparative studies focused on pathogenesis. Additionally, the fact that neither of these species are major pathogens will permit us to tease out virulence factors from genome comparisons.

Pilot shotgun sequencing data further support the value of *A. fischerianus* for comparative purposes. Comparison of contigs at 4X sequence coverage to *A. fumigatus* revealed sequence identities ranging from no hits for about 12 % of the contigs with the others ranging from 54% to 98% identity. An overall similarity is estimated at about 90%. This relatedness allows confident assignment of orthology between genes and regions in the two genomes, which is essential for comparative annotation.

Fig 2



A. fischerianus

*A. fischerianus (*a teleomorph of *Neosartorya fischeri),* a very close relative to *A. fumigatus*, is found in soil, and its spores are found in agricultural products. *A. fischerianus* can cause keratitis and possibly pulmonary aspergillosis in transplant patients, but is an extremely rare pathogen. Its inadequacy as a pathogen is interesting in light of its close evolutionary relationship to *A. fumigatus*, and so comparison of the *A. fischerianus* and *A. fumigatus* genomes should yield significant clues regarding *A. fumigatus* virulence and epidemiology. In addition, *A. fischerianus* has a known sexual cycle, and elucidation of *A. fumigatus* sexual reproduction by comparative genomics would be of immense value to the *Aspergillus* research community. This would greatly advance *A. fumigatus* as an experimental system by facilitating genetic studies. *A. fischerianus* has an active research community that will also benefit from the completed genome sequence. The type culture of *A. fischerianus* (NRRL 181) will be sequenced, using single spore subculture for genomic DNA extraction.

A. clavatus

A. clavatus is found in soils and animal manure, and is only rarely pathogenic, although potently allergenic. Interestingly, *A. clavatus* is located in the phylogenetic tree between two of the major pathogenic *Aspergilli, A. terreus* and *A. fumigatus*. As a result, *A. clavatus* is extremely well suited for comparative genomics studies designed to uncover *Aspergillus* virulence determinants. One non-pathogenic Aspergillus genome, *A. oryzae*, has been completed. Adding a second non-pathogenic species will greatly enhance the power of the comparison. Differences from the pathogenic *Aspergilli* shared by the non-pathogenic *Aspergilli, A. clavatus* and *A. oryzae*, will be excellent candidate virulence factors. The type culture of *A. clavatus* (NRRL 1) will be sequenced using single spore subculture for genomic DNA extractions.

Comparative Annotation

An immediate benefit of having genome sequences for *A. terreus*, *A. fischerianus*, and *A. clavatus*, will be the ability to have a highly accurate annotation for these genomes, as well as dramatically improving the annotation of *Aspergillus fumigatus*. The need for this is tremendous; at present roughly half of all predicted genes in the *A. fumigatus* sequence have no known function and lack homology to previous sequences. A comparative approach to annotation involving alignment of sequences from related species has been validated through the recent comparative annotation of *Saccharomyces cerevisiae* as well as the current drive to use sequence comparison to identify conserved functional elements in the human genome through comparison of the sequence with additional mammalian sequences. In the case of yeast, alignment of carefully selected related species enabled the most significant improvement of the yeast annotation since the release of the sequence ((Cliften 2003, Kellis 2003). In fact, annotations to nearly 15% of all genes in *S. cerevisiae* were altered based on these studies, including elimination of falsely identified genes, recognition of new genes, and corrections to the boundaries of many others. This contribution is especially significant given the amount of experimental data and manual curation supporting the *S. cerevisiae* annotation. In addition, comparative genomics data will facilitate development of species level diagnostics.

Depth of Community Interest, Community interactions

The Aspergillus research community is highly organized and cooperative, and the three organisms named in this whitepaper have been selected through an extensive public process over a number of years. In 2003 at an Aspergillus genomics workshop David Denning organized as part of the International Fungal Genetics meeting, Aspergillus researchers considered which new species would be most valuable sequencing targets and endorsed these species. A subsequent meeting in Copenhagen organized by Michelle Momany with 160 participants further considered this question and supported this choice. [*Aspergillus sydowi* was also discussed because of its close relationship to A. nidulans and its pathogenicity to sea corals]. Importantly, this same group is currently involved in an International collaboration involving many dozens of labs analyzing the genomes of *A. fumigatus, A. nidulans*, and *A. oryzae*. The lessons learned from studying these genomes, both individually and through comparative analysis, have refocused the community's interest on obtaining new genome sequences from *Aspergillus* species at appropriate evolutionary distances for the necessary studies. Moreover, the broader fungal research community has also been advocating for sequencing these same Aspergillus species. For example, at the initiation of the Fungal Genome Initiative in 2001, *A. terreus* was identified by the Steering Committee as one of the most important fungal targets for sequencing and was included in the very first FGI white paper submitted in 2002. It, along with *A. clavatus* and *A. fischerianus* was named again in the second FGI white paper submitted in 2003.

The highly interactive nature of the Aspergillus community that has made the joint analysis of the *A. oryzae, A. fumigatus*, and *A. nidulans* genomes so successful will again be harnessed to ensure that the sequencing and comparative analysis of these new sequences will also be successful. A number of Aspergillus researchers have directly contributed to this proposal and will constitute an Oversight Committee for the project. These include:

Michael J. Anderson (University of Manchester) (Aspergillus genomes curator, CADRE)
Michelle Momany (University of Georgia) (Chairman, Aspergillus Genome Policy Research Committee)
Gregory May (MD Anderson Cancer Center)
Thomas J. Walsh (National Cancer Institute)
David W. Denning (University of Manchester)

The resources developed to share existing data and the expanding analyses of the *A. nidulans, A. fumigatus*, and *A. oryzae* genomes will be used for these new genome sequences and ensure that the sequence, annotations, and analyses are accessible. For example, the Web sites already in place at TIGR, the Broad Institute and the University of Manchester currently serve Aspergillus genome sequences and the results of individual researcher's analysis. Further, the Central Aspergillus Data Repository at the University of Manchester (http://www.cadre.man.ac.uk) represents a centralized resource for Aspergillus genomic information, including curation of the sequence data and literature, and will incorporate data and findings from these new sequences, along with their existing presentations of data for *A. fumigatus*, *A. nidulans*, *A. oryzae*, and *A. flavus*.

DNA sequencing, assembly, and closure.

For each genome, paired-end sequence reads from multiple shotgun libraries prepared in different vector types with a variety of insert sizes will be produced. Test data from each library will be obtained separately and analyzed prior to approval for production sequencing. Genome data will be assembled using the assembly programs in production use at each of the MSCs. Genome assembly is an active area of research at each institution and improvements in the algorithms are implemented on an ongoing basis. The genomes of *A. terreus* and *A. clavatus* will be finished according to the standard protocols in place at each MSC. After completion of the shotgun assembly, *A. fischerianus* will receive additional automated targeted sequencing to close gaps and improve regions of low quality.

Specifically, we will:

- Generate deep draft genome assemblies (10X) and automated annotations for *A. terreus*, *A. clavatus*, and *A. fischerianus* (incorporating existing data for *A. fischerianus*).
- Perform additional automated sequencing for *A. fischerianus*, targeted to cover spanned gaps and regions of low quality in the assembly.
- Perform genome closure and finishing on *A. terreus* and *A. clavatus*.

Data release

In accordance with the NIAID's principles regarding data release, we will publicly release all data generated under this contract as rapidly as possible. As required by our contract, NIAID will be provided with a 21-45 calendar-day period to review and comment upon all data prior to its public release.

Chromatogram Files: Unless otherwise directed by NIAID, we will submit all sequences and trace files (chromatograms) generated under this proposal to the Trace Archive at NCBI on a no less than weekly basis. These data will also include information on templates, vectors, and quality values for each sequence.

Genome Assemblies: Genome assemblies will be made available via GenBank and the MSCs' websites, after internal and community validation. Assuming no significant errors are detected during the validation process, assemblies will be released within 45 calendar days of being generated.

Genome Annotation: Automated annotation data will be made available via GenBank and our web sites after internal and community validation. Assuming no significant errors are detected during the validation process, annotation data will be released within 45 calendar days of being generated.

Longterm curatorship: The assemblies and other pertinent data will be provided to the University of Manchester CADRE team (funded by the Wellcome Trust) for display and public access, along with other sequenced Aspergillus genomes, in an Ensembl-based format.

Ultimate value of the sequence

Like all genomic sequence, the data become a resource available for all time. As improvements in antifungal prophylaxis and prevention strategies emerge, *A. terreus* will remain one of the problem organisms, because of its ubiquitous nature in hospital water supplies and its high mortality rate. This sequence will provide a key tool for dealing with the problem.

Correct gene calling and understanding of gene function and expression are key objectives of genome sequencing. Given that about half of the genes found in *A. fumigatus* are hypotheticals and most new to science, a comparative approach to annotation becomes critical. Comparisons of secondary metabolic pathways and common allergic proteins, as well as virulence will be important to the field, and the former will greatly facilitate better bioinformatic prediction programs for secondary metabolic pathways, a key need.