Aspergillus flavus Genome Sequence: Initial Analysis

G. A. Payne[1], B. L. Pritchard[1], J. Yu[2], W. C. Nierman[3], R. A. Dean[1], D. Bhatnagar[2] and T. E. Cleveland[2]

[1] Center for Integrated Fungal Research, North Carolina State University

[2] USDA/ARS/SRRC [3] The Institute for Genomic Research

USDA NRI Knowledge For Tomorrow's Solutions
National Research Initiative Competitive Grants Program

OVERVIEW

Aspergillus flavus is a widely distributed filamentous fungus that normally occurs as a saprophyte in the soil or on decaying organic matter. It is pathogenic to both plants and animals and produces aflatoxin, one of the most toxic and carcinogenic naturally occurring compounds. *A. flavus* is also the second leading cause of aspergillosis in humans. A whole genome sequencing project funded by the USDA/NRI and USDA/ARS and conducted at TIGR is nearing completion. Preliminary studies of the 36.2 Mb draft sequence, which consists of 29 scaffolds representing 1407 contigs, indicate that the genome contains in the region of 11,760 genes. Thus the *A. flavus* genome appears to be larger than either *A. nidulans* or *A. fumigatus*. Aspergillus species are of interest in part because of the large range of clinically and industrially important secondary metabolites they synthesize. An initial scan of the *A. flavus* genome for proteins involved in the production of these secondary metabolites revealed the presence of 29 putative polyketide synthases and 24 putative non-ribosomal peptide synthetases. Manual annotation and analysis of the genome is being coordinated through North Carolina State University and will be made available at www.Aspergillusflavus.org.

PRELIMINARY GENOME STATISTICS

Several preliminary analyses have been run on the draft sequence of the A. flavus genome including gene prediction using the *ab initio* gene prediction programs GlimmerM[1] and SNAP[2]. In both these cases the gene finder was first trained using data from known Aspergillus genes. Based on supporting EST evidence and evaluation using a test set of genes, SNAP was found to be the more accurate prediction method and the following data refers to the set of genes predicted by SNAP.

Gene number	11,760
% GC content	49
Average gene length	1389 bp
Average coding length	1273 bp
Average number of exons per gene	2.59
Average intron length	71 bp
Number of genes supported by EST evidence (EST information from TIGR A. flavus gene index)	6705

Table 1.
Statistics for the A.flavus genome

Comparison of this preliminary *A. flavus* data with other members of the Aspergillus genus find that it possesses a larger genome than either *A.nidulans* or *A. fumigatus* and indicates that as expected it has more in common with *A. oryzae* than the other two aspergilli.

Species	Chromosomes	Size (Mb)	Gene number
A. fumigatus	8	29.2	10,000
A. nidulans	8	30.1	9,500
A. oryzae	8	37.6	13,572
A. flavus	8	36.2	11,760

Table 2.
How *A. flavus* compares with other members of the Aspergillus genus.

A preliminary scan of the *A. flavus* genome for genes that may be important in the production of secondary metabolites identified 29 putative polyketide synthases, 24 putative non-ribosomal peptide synthetases and more than 135 cytochrome P450 genes.

Species	Polyketide synthases	Non-ribosomal peptide synthetases	Cytochrome P450s
A. fumigatus	14	13	65
A. nidulans	28	14	102
A.oryzae	30	17	151
A. flavus	29	24	>135

Table 3.
Identification of genes involved in the production of secondary metabolites.

EXPECTED OUTCOME OF SEQUENCING PROJECT

An available complete genome sequence of *A. flavus* will greatly facilitate progress in understanding the ecology and evolutionary biology of the fungus, the regulatory networks controlling fungal development and primary and secondary metabolism.

There are four areas in which significant information will likely first emerge:

1) understanding of the evolution of secondary metabolism, its regulatory elements and its linkage to fungal development

- A direct comparison between *A. flavus* and *A. nidulans* should reveal conserved and diverged regulatory controls in secondary metabolic pathways, particularly aflatoxin and sterigmatocystin pathways. This information may provide clues as to the ecological significance of these compounds to the fungi.

2) factors in Aspergillus spp. responsible for human and plant pathogenesis

- A direct comparison among *A.flavus, A. fumigatus* and *A. nidulans*, a plant and animal pathogen, an animal pathogen and a saprophyte respectively, will likely provide insight into what constitutes a pathogenicity gene, and if different genes are required for plant and animal pathogenicity.

3) understanding of the phylogenetic relationships and evolution in the genus Aspergillus, and in particular the significance and evolution of gene clusters

- More species have been sequenced in the genus Aspergillus than for any other fungal genus, providing us with more power to address the evolution of the genomes of these fungi. As an example, the complete genome sequence of *A. flavus* will allow comparison of the evolution of the aflatoxin/sterigmatocystin cluster in relation to other genes within the Aspergilli. These studies can extend to examining the evolution of pathogenicity genes and genes important for saprophytic growth.

4) Insight into the way in which environmental pressure affects genome evolution.

- A direct comparison between *A.flavus* and *A. oryzae* may give insight into the way in which environmental factors affect genome evolution. Genetic and phylogenetic evidence suggests that *A. oryzae* is a domesticated biotype of *A.flavus*. Over the last few thousand years *A.oryzae* has undergone selection for production of advantageous secondary metabolites which has likely lead to differences in gene organization, gene family expansion and gene regulation between the 'domesticated' *A.oryzae* genome and the 'wild-type' *A. flavus* genome.

REFERENCES

1. **A.L. Delcher, D. Harmon, S. Kasif, O. White**, and **S.L. Salzberg.** 1999. Nuc. Acids Res. **27**(23): 4636-4641.
2. **Korf, I.** 2004. BMC Bioinformatics **5**:59
3. **Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S.** 2002. Genome Res. 2002 **12(10)**:1599-610.
4. **Kent, WJ.** 2002 Genome Res. **12**(4):656-64
5. **Lowe, TM and Eddy, SR.** 1997 Nuc. Acids Res. **25**:955-964

THE *ASPERGILLUS FLAVUS* GENOME BROWSER

A preliminary *Aspergillus flavus* genome browser has been set up using Gbrowse[3], the generic genome browser developed as part of the GMOD project. The browser currently contains information on 29 scaffolds that formed a 5X draft version of the genome, and the annotations resulting from several bioinformatic analyses that have been run on these scaffolds. Annotations currently available in the *A. flavus* genome browser include:

- Gene predictions from the gene-finding programs GlimmerM and SNAP.
- EST sequences aligned to the genome using BLAT[4].
- TIGR's *A. flavus* gene indices aligned to the genome using BLAT.
- tRNA genes predicted by tRNAscan-SE[5].
- BLAST matches to the *A. fumigatus* and *A. nidulans* genomes.
- BLAST matches to known Aspergillus proteins
- BLAST matches from the NCBI's nr database to the A. flavus predicted genes
- The location of restriction enzyme cleavage sites and potential transcription factor binding sites.

The genome browser provides an overview of each scaffold which can be customized by the user to show the locations of all the feature types they are interested in (Figure 1).

The browser also allows a user to search for a specific region of the genome or a particular feature of interest, access the annotation associated with this feature, see the relationship between different features and search for other related features as well as retrieve the underlying genomic DNA sequence.

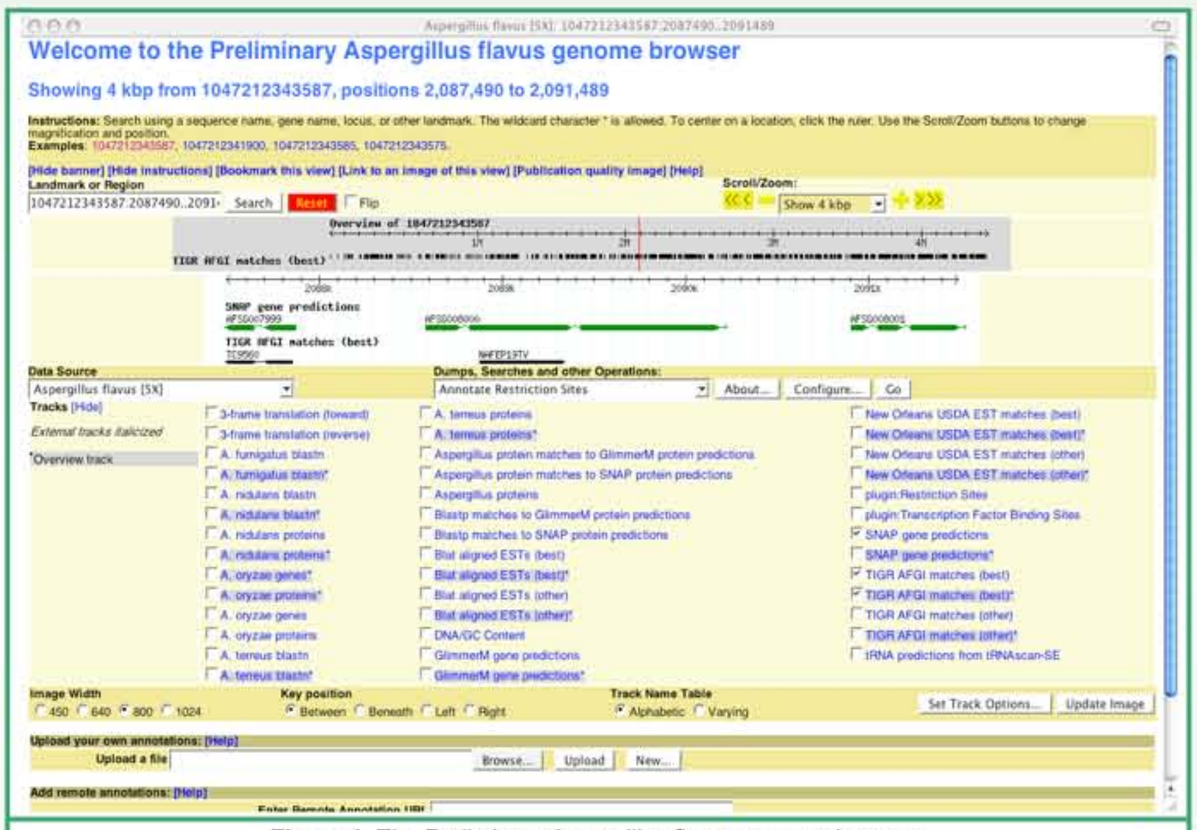

Figure 1. The Preliminary Aspergillus flavus genome browser.

DATA RELEASE

Date presentation

Data will be released from TIGR through their Comprehensive Microbial Resource (CMR see www.TIGR.org). Use of the CMR will allow analysis of gene function, biological role, placement into gene families, three-dimensional structure, links to other databases, and phylogenetic comparisons. *A. flavus* sequence and annotation also will be deposited in GenBank and at www.Aspergillusflavus.org, where an *A. flavus* genome browser will also be available. Manual annotation will be performed by the *A. flavus* research community and coordinated through North Carolina State University.

Community participation.

To ensure a close linkage with the Aspergillus flavus community and with the larger fungal community we have established a nine-member Steering Committee, which will aid in the oversight of the project and facilitate the dispersal of information gained from this study.

The members of this committee are:

- Gary Payne and Ralph Dean, North Carolina State University
- William Nierman, Vice President of TIGR
- Nancy P. Keller, University of Wisconsin
- Joan Bennett, Tulane University
- Gregory May, Baylor Medical School
- Charles Woloshuk, Purdue University
- Heather Wilkinson, Texas A&M University
- Jiujiang Yu, USDA/ARS/SRRC. Jiujiang Yu is a member of the International Aspergillus Genomics Steering Committee representing the A. flavus community.

Community Annotation.

Community annotation will be facilitated through the use of GBrowse as the genome browser for the *A. flavus* genome browser. In GBrowse different sets of annotations on the genome appear as separate tracks. This allows members of the community to create their own tracks based on the features they wish to annotate. These tracks may then be uploaded via DAS to the Genome browser to be viewed in the context of the complete genome sequence either by just the producer of the track or by the entire community. Following curation by the *A. flavus* annotation team at North Carolina State University these tracks may be incorporated into the core genome browser. Alternatively individual requests for genome features to be altered/annotated may be submitted to the annotation team via the A.flavus website www.Aspergillusflavus.org.

ACKNOWLEGEMENTS

This project was supported by the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant number # 2004-35600-14172.