

Minutes of the *Aspergillus fumigatus* genome annotation meeting, 5th to 6th May, 2003

Attendees:

Michael J Anderson (MJA), University of Manchester, UK
David Archer, University of Nottingham, UK
Kiyoshi Asai, National Institute of Advanced Industrial Science and Technology, Japan
Joan Bennett (JB), Tulane University, USA
Mark X Caddick, University of Liverpool, UK
David W Denning (DWD), University of Manchester, UK
Dennis Dixon (DD), National Institute of Allergy and Infectious Diseases, USA
Rory Duncan, National Institute of Allergy and Infectious Diseases, USA
Tamara Feldblyum, The Institute for Genomic Research, USA
Reinhard Fischer, University of Marburg, Germany
James Galagan (JG), Whitehead Institute for Biomedical Research, USA
Gustavo Goldman, University of São Paulo, Brazil
Neil Hall (NH), The Wellcome Trust Sanger Institute, UK
Linda Hannick, The Institute for Genomic Research, USA
Hoda M Khouri, The Institute for Genomic Research, USA
Jean-Paul Latgé (JPL), Pasteur Institute, France
Masayuki Machida (MM), National Institute of Advanced Industrial Science and Technology, Japan
Greg May, MD Anderson Cancer Center, USA
Bruce Miller, University of Idaho, USA
Michelle Momany, University of Georgia, USA
Ron Morris (RM), University of Medicine and Dentistry of New Jersey, USA
Karen Nelson, The Institute for Genomic Research, USA
William C Nierman (WN), The Institute for Genomic Research, USA
Stephen A Osmani (SO), Ohio State University, USA
Arnab Pain, Wellcome Trust Sanger Institute, UK
Ian Paulsen, The Institute for Genomic Research, USA
Geoff Robson, University of Manchester, UK
Steven Salzberg (SS), The Institute for Genomic Research, USA
Geoff Turner, University of Sheffield, UK
Owen R White (OW), The Institute for Genomic Research, USA
Jennifer Wortman, The Institute for Genomic Research, USA
Jiujiang Yu, United States Department of Agriculture, USA

Day One:

8.30 to 10:00 AM

Chair: William Nierman

WM welcomed all the attendees and they introduced themselves.

A. fumigatus (Af) sequencing project update

Tamara Feldblyum discussed progress at TIGR with finishing the *Af* genome. Six scaffolds are ready for annotation and two will be ready in two weeks. One scaffold, which contains the ribosomal repeat, is not in good shape. Ten telomeres have been linked to scaffolds and one centromere is contained within a BAC clone - this centromere will be sequenced if possible. One

chromosome has been joined together by linking clones and another is complete except for the centromere.

Neil Hall (NH) discussed progress at Sanger. They were given five scaffolds to finish, of which two are now continuous, one contains two contigs and one three contigs. They have been unable to detect telomeric or centromeric sequences in their scaffolds.

A. nidulans (An) sequencing project update

James Galagan (JG) discussed the whole genome shotgun (wgs) assembly of *An* that has been generated as part of the Fungal Genome Initiative. Four insert libraries have been used: 4 and 10 kb plasmids, 40 kb fosmids and 110 kb BACs. The wgs includes sequence from Monsanto. There are 89 scaffolds and 248 contigs, of which 95 % have been anchored to the genetic map. The whole of linkage group VIII is contained within one scaffold. The discrepancies between the genetic map and the assembly will be checked. Automatic annotation, as was used for *Neurospora crassa* (*Nc*) and *Magnaporthe grisea* (*Mg*), will be done in one month. The annotation pipeline involves three gene prediction programs and ESTs will be used to validate gene calls.

A. oryzae (Ao) sequencing project update

Masayuki Machida (MM) described this project: currently there are 50 sequencing gaps and 13 out of the 16 telomeres have been identified. Scaffolds have been mapped onto pulsed-field gels. Ten small contigs are unmapped. The contigs mapping to the largest chromosome add up to more than the expected size, but it is likely that the size as determined by pulsed-field gel electrophoresis, was an underestimate. However this chromosome currently contains three telomeres. An automatic annotation has been performed. The position of genes has been predicted using the program GeneDecoder, which has been trained using ESTs (~4000 unigenes).

Overview of the annotation process

Owen White (OW) introduced the subject by detailing what could be expected. He stressed that the annotation would be done automatically and that the data would be unstable, including the total number of genes. It is not practicable to 'freeze' the data as people need to get on with their jobs and many processes are of a serial nature. It was, however, possible to freeze the data by using versions, so that all the analyses could be done on the same version. The almost completed *Af* sequence could be annotated automatically and this annotation used as the starting point for analysis. However in the rush to publication, there would be no time to feed these analyses back into the database. There are implications for exchanging data between centres as they use different software and different sequence formats. This exchange will result in the loss of some data. OW illustrated how using different programs and individuals resulted in differences between sequencing centres by comparing a manually annotated 300 kb region of the *Arabidopsis* genome. For instance, Kazusa predicted 84 genes with 446 exons, while TIGR predicted 83 genes (which included 2 pseudogenes) with 455 exons. The most common discrepancies between the centres were at the 5' end and missing exons. However, 380 exons (~85 %) were identical between the three centres.

10:15 AM to 11:45 AM

Chair: William Nierman

Aspergillus Biology

Aspergillus disease and allergenicity

David Denning (DWD) gave an overview of the multiple forms of *Aspergillus* disease, which reflect variations in the interaction between the fungus and the host. He illustrated the relationship between immune function and type of disease and how the frequency of disease has increased over the years. Invasive aspergillosis has a 50 % mortality rate with dissemination to the brain occurring in 40 % of bone marrow transplant patients. Allergic bronchopulmonary aspergillosis occurs in 15 to 25 % of cystic fibrosis patients and is characterised by very high IgE titres and thickened airways. Twenty-two IgE-binding antigens have been recognised, which tend to be of a hyphal rather than conidial nature. In answer to a question, DWD stated that there is some variation between disease type and the species of *Aspergillus* involved: for instance, *A. fumigatus* disseminates more readily to the brain; *A. fumigatus* and *A. niger* are associated with aspergillomas and *A. niger* with infections of the ear.

A. fumigatus biology and pathogenicity

Jean-Paul Latgé (JPL) gave an overview of the biology of *A. fumigatus*, stating that it is a saprophyte and is the main fungus found in compost heaps. It is haploid with no detectable sexual stage and can grow up to 55 °C. It is a single non-recombining species and there is no clustering of clinical isolates into clades. Alveolar macrophages can be used as a cellular model for pathogenicity and various animal models have been developed, though these are not realistic of the clinical situation. Using mixed infections with several strains can be a more sensitive means to detect variation in pathogenicity by calculating the fungal burden of each strain recovered from tissue. Putative molecules with a role in pathogenicity include adhesins, pigments, toxins and various enzymes, but to date, no single pathogenicity factor has been identified in this species. He stated that there was a need to know what proteins are expressed in the lung. JPL demonstrated that swollen conidia are killed after phagocytosis by normal macrophages using reactive oxygen species. Finally he showed how the prevalence of *Aspergillus* species in the environment is reflected in the frequency of their isolation from patients.

A. oryzae processes and gene products of interest

Masayuki Machida (MM) overviewed the industrial importance of *A. oryzae*. Various classes of hydrolytic enzymes are important for the breakdown of soya beans, including glycosidases, lipases and proteases. He stated that they were interested in the regulators of the genes coding for these proteins and that they have identified the *cis*-acting promoter sequences upstream of these genes. It has been shown that *A. oryzae* is unable to synthesise aflatoxin because the gene for the transcriptional regulator AfIR contains a mutation and that one of biosynthetic genes is missing. Other areas of interest in the biology of *Ao* are the metabolic pathways important for fermentation and the secretory pathway.

1:00 to 3:00 PM

Chair: Owen White

Gene finding - overview of the programs and their performance

Steven Salzberg (SS) introduced the various gene-finding programs used by TIGR. Gene-finding programs locate the protein-coding region of the gene and always work better when trained on the

correct genome. GlimmerM, which uses an interpolated Markov Model algorithm, is now open-source; Exonomy, which uses a Hidden Markov Model algorithm, is currently not very user-friendly and requires a C++ expert; genesplicer is used for locating introns; and finally The Combiner utilises the output of six gene-finding programs, plus similarity hits to protein sequences and cDNAs. In an analysis of *Arabidopsis* genes, the Combiner predicted 78 % of the genes and 93 % of the exons correctly versus values of 31 % to 45 % for genes and 61 % to 79 % for exons with individual gene-finding programs. Finally, SS presented some data for *A. fumigatus*: when GlimmerM was trained with an *A. fumigatus* dataset, it predicted 48 % of the genes and 61 % of the exons correctly (using the same dataset), whereas when GlimmerM was trained with a larger *Aspergillus* dataset, it predicted only 35 % of the genes correctly.

The TIGR pipeline for automated structural annotation

Jennifer Wortman described how the Combiner package automatically predicts the initial gene set. The sequence is initially fed through a repeat masker and then gene predictions, EST alignments and protein alignments are performed before an optimal prediction is made. The gene prediction programs include GlimmerM, Phat and tRNASCAN, as well as splice site prediction programs. Sequence similarities are performed using AAT against protein databases and nucleotide databases (fungal gene indices, *Af* ESTs and other fungal ESTs). The evidence is weighted in the following order (highest first): match to a full-length cDNA, match to an EST and match to a protein. There is a viewer (Annotation station) for looking at the evidence.

The *Af* pilot sequence and a demonstration of ARTEMIS

Neil Hall described the 922 kb pilot sequence, which has been annotated manually by the Sanger team. Some statistics for this sequence are: 51 % G+C, 361 genes with 82 % containing at least one intron, average length of a gene: 1.4 kb with one gene every 2.6 kb, and 54 % of the genome is coding. He demonstrated that there was a degree of synteny with linkage group VIII of the *An* genetic map. The Sanger pipeline includes repeat masking, using BLASTN to find rDNA, and running tRNASCAN and gene-finding programs. These programs include GlimmerM, Genefinder and Phat. NH then demonstrated ARTEMIS and the Artemis Comparison Tool (ACT), which enables two or more genomes to be compared within an ARTEMIS environment. OW made the point that the tools being demonstrated were beyond the scope of most people at the meeting and that these demonstrations were illustrations of the annotation process.

3:15 to 6:00 PM

Chair: Owen White

Using the MANATEE viewer for functional annotation

Linda Hannick demonstrated the MANATEE viewer, which is used by TIGR annotators to curate genomes. A Web browser is used to access a genome database, which contains all the required pre-computed analyses, such as similarity searches: BLAST, SignalP, Pfam hits, etc. These analyses will have been generated by the Eukaryotic Genome Control (EGC) pipeline and loaded into a relational database. The central part of MANATEE is the gene curation page, where the evidence can be viewed as alignments or graphically. It is possible to look at the actual DNA or protein sequences and move left/right to the next gene. Gene Ontology (GO) suggestions will have been predicted automatically. It is possible to link to members of a paralogous family, which can be selected by Interpro domain or size and a selection of members of a family can be annotated in bulk or one-by-one - this is very useful to ensure consistency.

Naming conventions and Gene Ontology

Jennifer Wortman introduced the process of naming genes and proteins, which is done either by annotators at sequencing centres and is based on the literature and sequence similarities or by individuals after doing experiments. Automated naming can be done based on sequence similarity and domain hits, but can be victim to transitive annotation where a product name is transferred through several rounds of automated annotation. She described GO where a controlled vocabulary is used to describe gene products across organisms. Molecular function, biological process and cellular component are covered. The use of a controlled vocabulary facilitates uniform queries across the annotation of any genome. The EGC pipeline includes AutoGO where automated predictions are based on similar gene products and common Interpro domains. She emphasised that carrying out GO curation on an entire genome would be extremely time consuming, because the vocabulary is complex and the literature has to be reviewed. The community would have to address how to deal with fungal specific terms that had not yet been defined. It was generally agreed that a representation would be made to the GO consortium to identify new terms, relevant for filamentous fungi.

Neurospora genome annotation and writing the paper

James Galagan mentioned that manual annotation of two chromosomes has been performed by MIPS but that GO was not used. The entire genome was annotated automatically. They adopted an inclusive approach for the paper because some people submitted their own independent analyses. A steering committee was involved, biological subjects of interest were defined and the relevant people contacted. The process was co-ordinated by email, phone calls and some meetings at the Whitehead.

Comparative genomics

DWD introduced this subject (following a pre-meeting) by stating that the timelines for the three genomes under consideration (*Af*, *An* and *Ao*) were almost synchronous and it would be possible to analyse the genomes together. He proposed that at least three manuscripts could be prepared, one of which could be a comparative genomics paper.

SS described the MUMmer tool for carrying out genome alignments, which consists of NUCmer for nucleotide alignments and PROmer for amino acid alignments and a graphical viewer. The MUMmer tool looks for unique matches only between two genomes.

JG described how having three nearly complete genomes could provide an opportunity to improve the accuracy of the manual annotation, enable the study of genome evolution and permit the identification of conserved regions (such as non-coding RNAs and *cis*-acting sites). He emphasised these points using a few published examples, including a recent *Saccharomyces* comparative genomics paper where the comparison of the *S. cerevisiae* (*Sc*) genome with three other genomes enabled the number of 'real' ORFs to be defined.

Day Two:

8.30 A.M. to 1.00 PM

Chair: David Denning

Topics that will be covered by Sanger and TIGR

Neil Hall introduced which topics would be covered by the sequencing centres for the *Af* genome paper. It would depend of course on what was found. They have useful tools for metabolic

reconstruction. He personally would be interested in the sub-telomeric regions and evolutionary aspects, such as evidence of horizontal gene transfer. Bill Nierman stated they would cover basic genome issues, such as gene content, the structure of telomeres and centromeres, as well as the specialisms of team members: metabolism (Karen Nelson); membrane transporters (Ian Paulsen); DNA repair and metabolism (Jonathon Eisen) and comparative genomics (Jonathon Eisen). Ian Paulsen discussed membrane transporters in more detail: The transporters from 107 genomes have been classified by family and generalised substrate. The number of transporters is partly proportional to genome size, though soil organisms tend to have more than normal while intracellular parasites have fewer.

10 min talks: what proteins I would like to analyze and write about as a reflection of my research interests

Ron Morris (RM): His research interests in *An* have included histones, the cell cycle and microtubule genetics including the study of nuclear distribution (*nud*) mutants. These mutants have nuclei that do not move and have defects in motor proteins, such as dynactin and dynein, which interact with microtubules and membraneous cargo. He was concerned that, as these proteins are very conserved, whether this would make an interesting story for any of the papers; he thought that it would be more interesting to focus on the differences, such as sex, secondary metabolism and differences in sensitivity to temperature.

Steve Osmani (SO): His research interests in *An* include the cell cycle and its relationship to development: for instance, how the cyclin-dependant protein kinases, PhoA and PhoB, integrate environmental signals with asexual versus sexual development. He is also interested in nuclear pore complex proteins and how the nuclear pore opens up during mitosis to permit entry of tubulin, which condenses in the nucleus to form the spindle - this phenomenon is not seen in yeasts or higher eukaryotes.

Greg May: considered *An* a superb model for the whole genus. He is interested in the genetic programme for the growth of *Af* in an animal, which represents a starved environment for the fungus: what are the differences with a compost heap? His interests include MAP kinase signalling and its role in growth and pathogenicity, and the mechanisms of drug resistance.

Michelle Momany: Her research interests include polar growth, septation and morphogenesis, looking at germ tube emergence - how does polar growth begin? - and branch emergence. Polar growth involves two steps: the establishment of polarity and its maintenance. Much is known about these processes in *Sc* and homologues can be identified in *An*.

Mark Caddick: His research interest is in the field of gene regulation, which permits an organism to survive in a wide range of environments. For global regulators like AreA (nitrogen metabolism), genomics will permit the definition of all the genes under its control. Differences in such regulators can exist between organisms; for instance, penicillin biosynthesis is under nitrogen regulation in *Penicillium*, but production of the precursors of penicillin are not under N regulation in *An*. Similarly, regulation of toxin production seems very variable. He made the observation that a pathway for RNA degradation is present in *An*, but not in *Af* or *Nc*.

Bruce Miller: His research interest is in asexual and sexual development in *An*. Pathway specific transcription factors have been defined: e.g. asexual: BrlA, AbaA, WetA; sexual: StcA and NsdD. Two factors are involved in both pathways: StuA and MedA. Proteins involved in sexual development have been identified in *Af* whenever they have been looked for.

Reinhard Fischer: His research interests include nuclear migration, which is highly conserved between organisms and where no difference in gene complement would be expected, and the target genes of development. In *An*, an α -1,3- glucanase expressed in Hülle cells breaks down cell-wall glucan to glucose, which is taken up by the ascogenous hyphae using a high-affinity hexose transporter. A discussion ensued on other structures, such as cleistothecia and sclerotia and the genes that might be responsible for their formation.

Geoff Robson: His research interests in *Af* include phospholipases and signalling, and apoptosis. Three phenotypic markers for apoptosis have been detected during stationary phase in *Af* and also after treatment with amphotericin B and H₂O₂. *Af* secretes phospholipase B and three phospholipase Cs. When *Af* is grown in the presence of lipid, it spreads more rapidly because of a change in branching. It is possible in this instance that the breakdown products of phospholipases might be affecting signalling

David Archer: His research interests include protein secretion in *A. niger* and secreted proteins like proteases, cellulases and amylases. *A. niger* has yeast and human forms of the GTPases involved in vesicle formation. The process of glycosylation is more like that in *Sc* than in mammals. Differences between fungal species include: calnexin (involved in protein folding), which is essential in *S. pombe*, but not in *Sc* and *A. niger*; the HAC transcription factor involved in the unfolded protein responses whose activation is different in *Sc* and *A. niger*.

Joan Bennett (JB) and Geoff Turner: Their research interests are in secondary metabolism. Secondary metabolites are low molecular weight natural products usually produced after active growth whose biological function is often unknown. The genes for their biosynthesis are usually in clusters and can code for large multi-domain proteins like non-ribosomal peptide synthetases and polyketide synthetases. Currently it is not possible to predict which amino acids are utilised by a peptide synthetase. The challenge will be to match up the secondary metabolites known to be produced by *Af* with the gene clusters identified.

Gustavo Goldman: His research interests are anti-fungal drug resistance in *Af* and DNA repair in *An*. Most current drugs target a step in ergosterol biosynthesis and it is already known that *Af* has two *ERG11* genes. One mechanism of resistance is increased expression of ABC and major facilitator transporters and it should be possible to identify members of these two classes of transporter that might be involved in drug efflux.

Jean-Paul Latgé: His research interest is in cell wall structure and its biosynthesis. Some components of the cell wall, such as α -1,3- glucan and galactomannan, are unique to *Aspergillus*, but currently the enzymes involved in their biosynthesis are unknown. There are differences between the structure of *Aspergillus* and *Sc* cell walls: *Af* has no α -1,6- glucan or linear α -1,6- mannan, though it has genes coding for homologues that are responsible for the biosynthesis of these polysaccharides in *Sc*.

David Denning: is interested in hormone receptors and insulin-like factors in *Aspergillus*. Corticosteroids have been shown to stimulate growth of *Af* and *A. flavus*, but not *An*. A corticosteroid receptor has been identified in *Candida albicans*. An insulin-like function and immunoreactivity has been found in *Af*.

1.00 to 4.00 PM

Chair: Neil Hall

Discussion: on the processes of data transfer, a confidentiality agreement, data release, comparative analyses, possible inclusion of the *A. niger* sequence data from DSM, additional datasets that would be useful, timelines and co-ordination issues.

The meeting concluded at 4:00 PM.