

Minutes of the April meeting of the *Aspergillus fumigatus* (Af) genome sequencing group

The University of Salamanca, Spain, 11 and 12th April 2002

Meeting attendees: (\*steering committee)

University of Manchester:

David Denning (DWD)\*

Michael Anderson (MA)\*

Jane Mabey

Peter Giles

The Wellcome Trust Sanger Institute:

Neil Hall (NH)\*

David Harris (DH)

John Woodward (JW)

Michael Quail (MQ)

Marie-Adele Rajandream

The Institute for Genomic Research (TIGR):

William Nierman (WN)\*

Tamara Feldblyum (TF)

Owen White (OW)

University of Salamanca:

Miguel Sánchez-Pérez (MS)\*

Fernando Leal

David Knowels

Complutense University:

Javier Arroyo

José Manuel Rodríguez Peña

Centro de Investigaciones Biológicas:

Miguel Angel Peñalva (MP)

Institut Pasteur:

Jean-Paul Latgé\*

National Institute of Allergy and Infectious Diseases:

Dennis Dixon\*

Apologies: Geoff Turner\*, University of Sheffield; Joan Bennett\*, Tulane University

Thursday 11<sup>th</sup> April

After an introductory talk from DWD about the medical and scientific importance of *Aspergillus fumigatus* and other *Aspergillus* species, progress reports were presented by the various centres.

1) Progress with shotgun sequencing (Spain): MS stated that the Spanish consortium had generated ~20,000 reads from a library sent to them by MQ. These reads had been sent to the Sanger.

2) The physical map and pilot project (Sanger): JW described the construction of the BAC libraries and their use in generating a *Pst*I fingerprint map. This map is currently

undergoing its final edit. He then described the pilot project which involved sequencing 28 BAC clones to make up a 850 kb contig, centred around the *niaD* gene. This contig maps to the smallest chromosome (1).

3) Whole genome shotgun sequencing (Sanger): DH described the Sanger contribution to the WGS. Seven production libraries were made with most reads generated from 2.2 – 2.5 kb and 2.5 - 4.0 kb libraries. He then described the assembly process and presented the statistics from their assemblies using only Sanger reads and both Sanger and TIGR reads. For instance, Sanger had generated ~ 240,000 reads of average length 540 bases. He finished off by illustrating some points using the largest assembled contig – for instance, the premature termination of the sequencing reactions at both ends.

4) Whole genome shotgun sequencing (TIGR): TF described the TIGR contribution to the WGS. Four production libraries were made (2 - 3 kb, 3 - 4 kb, 10 –12 kb and 50 kb). She then presented the statistics about the 10 x WGS, which includes Sanger traces (*e.g.*, 97 % of the bases are in contigs > 10 kb with the largest contig being 1,350,153 bp). She illustrated some of the process of checking the assembly with examples. For instance, reads containing the telomeric repeat mainly fell into one contig which had read pairs that linked to six other contigs. She finished off with an example of the automated annotation that they will put in place with the next assembly to be released into the public domain (localisation of ORFs and BLAST reports).

5) General discussion:

a) Centromeres: It was agreed that it might be possible to identify putative centromeric sequences by looking at the ends of contigs. DWD proposed that an attempt be made to obtain the complete sequence of chromosome 1. It was suggested that if a sub-centromeric or centromeric probe could be identified, then the BAC libraries could be screened to identify clones for sequencing. If it was felt necessary, a YAC library could be constructed which might contain clones that span a centromere.

b) ESTs: There was general agreement that ESTs were useful for confirming gene predictions. However, OW stated that full-length cDNA sequences were much more valuable. It was suggested that 1000 full-length cDNA sequences be generated and that these data could be published separately. The sequences would be required at the end of the finishing/closure phase of the project. It was thought that sequencing another strain (*i.e.*, AF10) would be more interesting as SNPs might be identified. MS is to establish if the Spanish consortium have sufficient funding for this project.

6) Steering committee meeting

Friday 12<sup>th</sup> April

6) Finishing/closure workshop:

a) Assembly: It was determined that the first task that needs to be done is for TIGR to generate a new assembly using the additional data from Sanger/Pasteur (BAC end pair sequences) and Spain (production library traces). This should result in larger groups/scaffolds and a smaller bin. The Sanger agreed to use this assembly. The scaffolds would be split evenly between the two centres. Both centres would verify

the assembly of repetitive sequences (TIGR (TF) to supply Sanger with the output of RepeatFinder), ensure that there was at least 2 x coverage of all bases and perform manual editing.

b) Sequencing gaps: It was thought likely that there would be sufficient clone coverage for each centre to be able to fill any gaps with their clones. The problem with exchanging libraries was that pBR-based libraries (as generated by TIGR) could not be used at the Sanger because of GMO regulations. It was a possibility that the modifications carried out on these plasmids might have disabled them sufficiently and so this will be looked into (WN and MQ). Sequencing gaps would be considered closed when covered by 2 different clones, chemistries or in both directions.

c) Physical gaps: It was proposed that some of these gaps would be joined by using multiplex PCR. Additional genomic DNA would probably be required for this and MA is to provide more. It was felt that a conference call might be needed at this point or alternatively NH and DH could go to TIGR in Jul/Aug in order to decide what is to be done and by whom. It was proposed that one way of tracking physical ends would be to exchange FastA files of the last 1000 bp. Centres would also have to exchange scaffolds when they were joined together.

d) Project management: OW is to see about setting up a website to enable the two centres to track which traces from the bin have been used. TF and DH are to exchange initial information and then to provide the contact names in their closure teams.

e) DWD proposed the following suggestions: As a last resort, the other centre could attempt to fill a sequencing gap; if one centre was progressing more quickly, they could move onto closing the physical gaps.

7) Annotation workshop: It was affirmed that the annotation would not start until the sequence is complete. Annotation standards have been decided between Sanger and TIGR for *Plasmodium falciparum* and both centres were open to using gene ontology (<http://www.geneontology.org/>)(GO). The centres would use XML as the file transfer format. A database needs to be set up to house the annotation – this was proposed to be at TIGR. A data use seminar was proposed where the biologists involved in writing the paper would be introduced to this database, its use and content. These biologists would have to be able to deal with GO and possibly have to develop new terms.

8) Experimental resources for the community: MA outlined the potential resources that could be provided for the community including minimal clone sets, one derived from the 10 – 12 kb plasmid library and one from the BAC libraries. MP suggested that a cDNA library would be useful. It was suggested that it would be possible to have a resource centre in the UK as well as in the US.

9) Proposed timelines: an Excel spread sheet was constructed with detailed timelines for each participant.

10) Manuscript: it was reported that the steering committee had agreed on producing one paper for Nature or Science and that a writing committee had been setup (WN, NH, DWD, MA)

Michael J. Anderson  
May, 2002.