Aspergillus Genomes and the Aspergillus Cloud

Jane E. Mabey Gilsenan1,* , Graham Atherton1, Jennifer Bartholomew1 ,
Peter F. Giles1, Teresa K. Attwood2,3, David W. Denning1 and Paul Bowyer1

1School of Medicine, The University Hospital of South Manchester (Wythenshawe), Manchester M23 9LT,
2Faculty of Life Sciences and 3School of Computer Science, The University of Manchester, Oxford Road,
Manchester M13 9PT, UK

Received September 14, 2008; Accepted October 18, 2008

ABSTRACT
Aspergillus Genomes is a public resource for viewing annotated genes predicted by various Aspergillus sequencing projects. It has arisen from the union of two significant resources: the Aspergillus/Aspergillosis website and the Central Aspergillus Data REpository (CADRE). The former has primarily served the medical community, providing information about Aspergillus and associated diseases to medics, patients and scientists; the latter has focused on the fungal genomic community, providing a central repository for sequences and annotation extracted from Aspergillus Genomes. By merging these databases, genomes benefit from extensive cross-linking with medical information to create a unique resource, spanning genomics and clinical aspects of the genus. Aspergillus Genomes is accessible from http://www.aspergillus-genomes.org.uk.

INTRODUCTION
An important consequence of the impact of global climate change, brought to the fore by recent natural disasters, has been a stimulation of interest in fungal ecology. In the aftermath of Hurricanes Katrina and Rita, for example, residential areas in New Orleans, having remained underwater for weeks, succumbed to rapid mould growth (1–3), facing residents with significant health hazards: exposure to high concentrations of some moulds (e.g. Aspergillus, Cladosporium and Trichoderma) can cause severe and sometimes life-threatening responses. Examination of samples taken from affected homes identified several Aspergillus species, Aspergillus niger being predominant amongst them. The concentration of this and other species reached levels generally associated with environmental health problems. The paucity of effective drugs available to treat Aspergillus infections, coupled with the speed of diagnosis required for successful treatment, raises concerns for residents and tradesmen who return to restore such areas without taking appropriate protective measures.

Generally, Aspergillus is a genus of fungus found worldwide: approximately 250 species descriptions have been published (4,5), several of which are of medical or industrial importance. Aspergillus fumigatus is the most common mould pathogen of humans, causing both life-threatening invasive disease of immuno-compromised patients and allergic disease in patients with atopic immune systems (6–8). Aspergillus nidulans, an occasional human pathogen, is a model organism that has contributed to our understanding of genetics, gene regulation and cellular biology (9,10), while A. niger (11,12) and A. oryzae (13) are both used in industrial processes. Aspergillus flavus is both a human and plant pathogen, being responsible for a disproportionate number of cutaneous and wound infections in man (14). Several other Aspergillus species are known to be significant allergens or to be responsible for mycotoxin production in stored food (15–17).

In light of their relevance to medicine and industry, and the desire to better understand this genus, the genomes of 10 Aspergilli have recently been sequenced, seven of which have been annotated in worldwide collaborative efforts. To be of maximum benefit to the community in general, this expanding pool of genomic data (the volume of which is likely to increase with the decreasing cost of sequencing) requires collation and long-term maintenance; more particularly, to benefit research into Aspergillosis and other illnesses caused by this genus, the accumulating genomic data needs to be cross-linked to the numerous Aspergillus-related medical reports now available. However, the standard model for genome databases [e.g. SGD (18)] is not appropriate for medically significant organisms, where the dissemination of information, such as patient care, drug resistance or toxins is equally important. Thus, Aspergillus Genomes was established in April 2008 to provide links between genomic and medical information and to provide access through one portal.

To whom correspondence should be addressed. Tel: +44 161 2915908; Fax: +44 161 2915804 Email: jane.gilsenan@manchester.ac.uk

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
SOURCE DATA AND METHODS

This resource is a joint undertaking by two teams that have, until now, provided online medical and genomic information for different groups within the Aspergillus community.

Medical information

The Aspergillus/Aspergillosis website (http://www.aspergillus.org.uk) was designed to serve the medical and translational research communities, including medical consultants, scientists, patients and their relatives. Offering a wealth of information about Aspergillus and the multiple diseases (known collectively as Aspergillosis) it can cause, one section provides advice and a discussion group for patients, while other areas provide medical and species images, educational materials, meeting reports and publications. As part of Aspergillus Genomes, this large collection of material offers a valuable repository from which we can begin to derive much-needed medical insights from genome analyses (Some sections require registration, please refer to supplement ‘Note: Access information’).

Genomic information

The Central Aspergillus Data Repository, CADRE (http://www.cadre-genomes.org.uk) (19), was developed primarily to serve the Aspergillus genomics community. Its principal role has been to manage genomic data and to offer web-based tools for analysis and visualization of genomic features. These tools offer simple displays for viewing annotation of predicted genes (e.g. function, GO terms, similarity matches) and of their protein products (e.g. family and domain similarity matches), as well as complex displays for viewing genes and other features (e.g. RNA-encoding genes, repeated sequences) in the context of an assembly.

Incorporation of this resource into Aspergillus Genomes provides access to seven annotated genomes, five of which are of particular importance to the Aspergillus genetics community (A. fumigatus Af293 and A1163, A. nidulans, A. oryzae and A. niger); of these, two are clinical isolates (A. fumigatus Af293 and A1163).

CONTENT OF CURRENT RELEASE

Aspergillus Genomes has been implemented using Ensembl version 22 (20), a suite devised for storing annotated eukaryotic genomes. The latest release (June 2008) contains information pertaining to seven genomes.

Aspergillus fumigatus (Af293) 2.2.1

Using whole-genome shotgun (WGS) sequencing (carried out by TIGR) and optical mapping, this project yielded an assembly of 19 supercontigs (~28.8 Mb) mapping onto eight chromosomes. Using automated gene predictions, the University of Manchester coordinated an international manual annotation project, and data for 9926 genes were released in 2005 (21). This genome has since been re-annotated as part of an additional international comparative project led by the J. Craig Venter Institute (JCVI) and the University of Manchester. Data for 9630 predicted protein-coding genes were publicly released in 2008 (22).

Aspergillus nidulans (FGSC A4) 5.3.1

WGS sequencing of this genome was undertaken by the Broad Institute and publicly released in March 2003. This project yielded an assembly of 17 supercontigs (~30 Mb) assigned to eight linkage groups, a group of unassigned fragments (84 contigs) and 9520 predicted protein-coding genes. Further automated and manual annotation carried out by TIGR (2005) yielded a total of 10 542 genes (23)—using publicly available linkage data (John Clutterbuck, http://www.gla.ac.uk/ibs/molgen/aspergillus/index.html) and in-house analyses, we have updated the original Broad assembly to reflect these findings. More contigs have now been mapped, with correct orientation, within each supercontig; the supercontigs have also been orientated correctly within each linkage group.

Aspergillus oryzae (RIB 40) 1.1.1

WGS sequencing of this genome was undertaken by a Japanese Consortium, led by the National Institute of Advanced Industrial Science and Technology (AIST). Using optical mapping, the project yielded an assembly of 22 supercontigs (~37.1 Mb) mapped to eight chromosomes and a group of unassigned fragments. Using automated annotation techniques, 12 074 protein-coding genes were predicted and publicly released in 2005 (24).

Aspergillus niger (CIB 513.88) 1.1.1

Using WGS sequencing and BAC walking (carried out by DSM Food Specialties), this project yielded 19 supercontigs (~33.9 Mb) mapping onto eight chromosomes. Using a base of automated gene predictions, DSM coordinated a European manual annotation project, predicting 14 086 protein-coding genes, and data were released in 2007 (25).

Aspergillus fumigatus (A1163) 1.1.1, A. clavatus (NRRL 1) 1.1.1 and Neosartorya fischeri (NRRL 181) 1.1.1

These genomes were sequenced using WGS sequencing (carried out by JCVI), as part of an international comparative project led by JCVI and the University of Manchester (2008) (22). The work on A. fumigatus yielded 55 scaffolds (~29 Mb), for which 9929 protein-coding genes were predicted; A. clavatus yielded 143 unassigned fragments (~27.8 Mb), for which 9121 protein-coding genes were automatically predicted; and N. fischeri yielded 976 unassigned fragments (~32.5 Mb), for which 10 406 protein-coding genes were automatically predicted.

DISPLAY AND SEARCH SOFTWARE

Several tools are provided (via CADRE) for viewing genomic data within Aspergillus Genomes. GeneView (Supplementary Figure S1) is the principal data-visualization tool, offering detailed information about a particular gene, including the public locus, the chromosomal location and a short description of the gene. An overview of each predicted transcript is also provided: this consists of
structure information, a list of database cross-references to similar sequences (e.g. SGD) and protein features [e.g. InterPro (26)], and GO terms that have been mapped to the gene. Images are also displayed of each transcript structure and of domains or family signatures mapped to the protein.

In addition to the information provided in GeneView, Transview provides the transcript sequence, which can be marked up to highlight both codons and the peptide sequence. ExonView provides the nucleotide sequence for each exon, as well as the upstream and downstream regions, untranslated regions and introns. ProtView displays further information about a particular protein and, in addition to the GeneView summary it provides the sequence in FASTA format, which can be marked up to highlight exons. ProtView also provides some predicted peptide statistics and information about any matches to family- or domain-based databases [e.g. Pfam (27), PRINTS (28)].

Four search tools, all accessible from the homepage, are available for Aspergillus Genomes: UniSearch, Google™ customized search, the Aspergillus Cloud and BLAST. UniSearch, the in-built Ensembl search tool, allows users to search public identifiers assigned to genes (e.g. AFUB_063690 and An14g07380) — it is found on the homepage and throughout the website. A familiar tool is provided in the form of a customized Google™ search engine, thus enabling free-text searches on all information pages and genomes within the resource. A more recent addition is ‘cloud’ searching, which is described in more detail below. To allow similarity searches, BLAST can be performed against in-house sequences (i.e. genomic sequence, predicted ORFs and proteins), the results of which are presented in the context of the assembly rather than as stand-alone alignments.

THE ASPERGILLUS CLOUD

The results of UniSearch or Google™ searches are provided in the form of lists of pages and documents matching the given term. Although useful in many cases, this type of output does not empower the user to access and search related data. To remedy this, we have introduced the Aspergillus Cloud, a facility that uses ‘cloud’ searching (developed by Quintura™) across the medical and genomic information embraced by Aspergillus Genomes. Cloud searching provides a visual means of searching related terms: when a search term is entered, this generates a group (or cloud) of related terms from context, as well as a list of hyperlinks to relevant pages/documents. Furthermore, by exploring the cloud (i.e. by moving a mouse over any term), search results change to reflect relationships with the new highlighted term.

We can demonstrate the Aspergillus Cloud using documented research on point mutations in the cyp51a gene and its correlation to rising triazole drug resistance. By simply entering ‘cyp51a’ into the search box (Figure 1), a number of related terms are displayed, as well as a number of hyperlinks to relevant pages. Several links provide access to pages that are contained within the medical section of Aspergillus Genomes and refer to point mutations within the gene and drug resistance; one provides access to the relevant gene within the genomic section of Aspergillus Genomes. Medical terms can also be used to explore the Aspergillus Cloud (Figure 2). By entering the term ‘Allergic Bronchopulmonary Aspergillosis’ (ABPA; an allergic reaction to Aspergillus spores), opportunities are given to explore articles written about the allergen to which ABPA patients are hypersensitive, as well as annotation provided for an ABPA-related gene.

The Aspergillus Cloud provides a powerful means of exploring Aspergillus data and finding medical reports linked to particular genes that would require several search attempts or would otherwise be missed using conventional search tools.

EXPLORING FEATURES OF INTEREST

The collation of Aspergillus data enables us to focus on and embellish features of medical interest. Aspergillus fumigatus, for example, possesses over 80 genes encoding allergen proteins and of these 22 have been cloned as IgE binding (29,30). We have documented these allergens within Aspergillus Genomes and provided links to annotated genes (http://www.aspergillus.org.uk/indexhome.htm?secure/sequence_info/genesofinterest/allergens.php). Another area of great interest is secondary metabolites. We have already extracted information on many secondary metabolites (http://www.aspergillus.org.uk/secure/metabolites/list_by_secret.php) and intend to link this data to annotated genes. We also have begun the process of highlighting secondary metabolite clusters, such as the gliotoxin biosynthetic cluster. Gliotoxin is a toxin produced by A. fumigatus and possesses many biological properties, including the ability to act as an antibacterial and antiviral agent as well as an immunosuppressant (31,32). We have provided a view of the gene cluster within the assembly and links to the appropriate annotation (Figure 3).

FUTURE DIRECTIONS

In the short term, we plan to add three more new genomes and to revise a fourth. The new genomes include A. flavus NRRL 3357 (TIGR, 2005), A. terreus (Broad Institute, 2006) and A. niger ATCC 1015 (DOE Joint Genome Institute, 2006); although these genomes have been sequenced with high coverage, only draft sequence and limited annotation are currently publicly available—we will add the sequence data to our collection for comparative work and add annotation as it becomes available. As for the latter, we are currently participating in Eurofungbase, a European project coordinated by Eurofung (http://www.eurofung.net), to manually re-annotate A. nidulans with the help of experts from 32 laboratories and 13 industrial partners. Our role is to house the corrected gene structures and annotation arising from this project, and to make the data publicly available: we are processing this information in-house, and will release it through CADRE and Aspergillus Genomes.
Figure 1. Using the Aspergillus Cloud to explore genes. Entering a gene name, such as ‘cyp51a’, into the search box at the top of the cloud search space (a) produces numerous related search terms immediately below and a list of relevant documents/records at the RHS. The document list provides an extract for each document enabling the user to assess the context of the search term found. By clicking on relevant links, the user can access both medical reports (b) and genomic annotation (c) for a given gene name.

Figure 2. Using the Aspergillus Cloud to explore medical information. Entering a medical condition, such as ABPA, into the search box (a), produces several groups of related terms within the search space below. By moving the mouse over each term, new clouds arise temporarily. Clicking on a term will add the associated clouds into the search space and change the focus of the search to include the term (b). In this example, we find a list of documents on the RHS that includes links to an ABPA-related gene (c) and to an article regarding an allergen that is particular to the allergic response of ABPA patients (d).
In the longer term, we will continue to expand the section dealing with areas of medical interest. Explicit links and clustering of such data enables rapid assimilation of essential information. In addition, to support in-house work and provide users with value-added data, we will expand our comparative analyses and include resultant data within Aspergillus Genomes. Access to similar data is currently facilitated by the Broad Aspergillus Comparative Database (http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiHome.html); however, in contrast with Aspergillus Genomes, which places much attention on the underlying data, the Broad site focuses on function. By providing up-to-date primary data, we hope to yield consistent secondary data, as well as relevant links to current medical information.

CONCLUSIONS

The Aspergilli garner interest owing to their medical and industrial significance, but changes in the environment are making the interest in this genus and its effects on the human condition more prominent. It is therefore timely to begin the process of marrying medical and genomic information. By maintaining a resource that accommodates both such data types, and by providing several means of viewing, searching and analysing the data, we hope to better serve the Aspergillus research community.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We wish to thank all those who have contributed to Aspergillus Genomes by their involvement in the sequencing and annotation projects. Interactions with the A. nidulans research community are facilitated by workshops organized by Eurofung (Eurofungbase project—EC 1274195).

FUNDING

The Fungal Research Trust.
