New resources for functional analysis of omics data for the genus *Aspergillus*

Benjamin M Nitsche1*, Jonathan Crabtree2, Gustavo C Cerqueira3, Vera Meyer1,4,5, Arthur FJ Ram1,4 and Jennifer R Wortman3

**Abstract**

**Background:** Detailed and comprehensive genome annotation can be considered a prerequisite for effective analysis and interpretation of omics data. As such, Gene Ontology (GO) annotation has become a well accepted framework for functional annotation. The genus *Aspergillus* comprises fungal species that are important model organisms, plant and human pathogens as well as industrial workhorses. However, GO annotation based on both computational predictions and extended manual curation has so far only been available for one of its species, namely *A. nidulans*.

**Results:** Based on protein homology, we mapped 97% of the 3,498 GO annotated *A. nidulans* genes to at least one of seven other *Aspergillus* species: *A. niger*, *A. fumigatus*, *A. flavus*, *A. clavatus*, *A. terreus*, *A. oryzae* and *Neosartorya fischeri*. GO annotation files compatible with diverse publicly available tools have been generated and deposited online. To further improve their accessibility, we developed a web application for GO enrichment analysis named FetGOat and integrated GO annotations for all *Aspergillus* species with public genome sequences. Both the annotation files and the web application FetGOat are accessible via the Broad Institute’s website (http://www.broadinstitute.org/fetgoat/index.html). To demonstrate the value of those new resources for functional analysis of omics data for the genus *Aspergillus*, we performed two case studies analyzing microarray data recently published for *A. nidulans*, *A. niger* and *A. oryzae*.

**Conclusions:** We mapped *A. nidulans* GO annotation to seven other *Aspergillus*. By depositing the newly mapped GO annotation online as well as integrating it into the web tool FetGOat, we provide new, valuable and easily accessible resources for omics data analysis and interpretation for the genus *Aspergillus*. Furthermore, we have given a general example of how a well annotated genome can help improving GO annotation of related species to subsequently facilitate the interpretation of omics data.

**Background**

Gene Ontology (GO) is a framework for functional annotation of gene products aiming to provide a unique vocabulary for living systems [1]. It comprises Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) ontologies. GO terms are organized as directed acyclic graphs (DAG) meaning that GO terms are connected as nodes by directed edges defining hierarchical parent-child relationships. As a consequence, the specificity of GO terms increases with increasing distance from their root node. Enrichment analysis of GO terms is a well accepted approach to dissecting omics data in a non-biased manner. It has been used in many studies to highlight major trends in genomic, transcriptomic or proteomic datasets and describe them with a controlled vocabulary [2-5]. If the frequency of specific GO terms in a list of genes or proteins is higher than expected by chance, it is likely that these enriched GO terms are related to the biological processes under investigation.

The genus *Aspergillus* covers a group of filamentous fungi that includes saprophytes, human and plant pathogens as well as species being exploited in biotechnology. Whereas *A. nidulans* has been comprehensively studied and used as model organism, *A. niger*, *A. oryzae* and *A. terreus* are important industrial workhorses for the...
production of various enzymes and organic acids. In medical research, *A. fumigatus* and *Neosartorya fischeri* are intensively studied because of their importance as allergens and pathogens of immunocompromised patients. The aflatoxin producing fungus *A. flavus* is well known to cause spoilage of a great variety of agricultural goods. With genome sequences publicly available for eight of its species, the genus *Aspergillus* provides an important group of related fungal species for comparative genomics [6]. The exceptional role of this genus in the genomics of filamentous fungi is further emphasized by a community sequencing project for comparative genomics [6]. The number of predicted transcripts were obtained from the Central Aspergillus Data REpository (CADRE) [40].

Table 1 Mapping of *A. nidulans* GO annotation

| Species       | Strain    | Predicted GO annotated (%) | GO terms |
|---------------|-----------|----------------------------|----------|
| *A. nidulans* | FGSC A4   | 10546                      | 3498 (33)| 5508     |
| *A. fumigatus*| AF2937    | 9846                       | 3443 (35)| 5445     |
| *A. fumigatus*| A1163     | 10109                      | 3450 (34)| 5446     |
| *A. flavus*   | NRRL 3357 | 13487                      | 3574 (26)| 5463     |
| *A. niger*    | CBS 513.88| 14366                      | 3540 (25)| 5430     |
| *A. oryzae*   | ATCC 1015 | 11200                      | 3487 (31)| 5412     |
| *A. terreus*  | NIH 2624  | 10402                      | 3414 (33)| 5406     |
| *N. fischeri* | NRRL 1    | 9379                       | 3403 (36)| 5449     |
| *N. fischeri* | NRRL 181  | 10728                      | 3543 (33)| 5445     |

Summary of mapping *A. nidulans* GO annotation to seven other Aspergilli. The number of predicted transcripts were obtained from the Central Aspergillus Data REPOSITORY (CADRE) [40].

In this work, we have generated a new central repository for functional analysis of omics data for the genus *Aspergillus* using GO annotation. Firstly, we extended the GO annotation of *A. nidulans* to all *Aspergillus* species with publicly available genome sequences and generated annotation files compatible with diverse publicly available tools for GO enrichment analysis. Secondly, we further improved the accessibility of the GO annotation for the genus *Aspergillus* by integrating it into a web tool for GO enrichment analysis and graph visualization named Fisher’s exact test Gene Ontology annotation tool (FetGOat). Finally, we performed two case studies to demonstrate the value and flexibility of the newly generated resources for functional analysis of omics data for the genus *Aspergillus*.

Results

Mapping of GO annotation

*A. nidulans* is the only *Aspergillus* species for which comprehensive GO annotation based on both computational prediction and extended manual curation of gene-specific literature is available [9]. It constitutes a valuable resource for GO enrichment analysis, which has proven to be a powerful tool for dissecting omics data, for example sets of differentially expressed genes. The GO annotation of *A. nidulans* available at the Aspergillus Genome Database (AspGD) [9] covers 33% (3,498) of its predicted transcripts and associates them with 3,340 GO terms. Including all parental nodes, the list of GO terms extends to 5,508 comprising 3,061 (55%) BP, 1,753 MF (32%) and 694 (13%) CC terms.

To extend this valuable resource to other species of its genus, we mapped the *A. nidulans* GO annotation to all *Aspergillus* strains with published genome sequences (see Table 1). Groups of orthologous and close paralogous proteins were compiled with the Sybil comparative analysis package [10], which applies a modified reciprocal best-hit approach comprising two clustering cycles. Roughly 89% (99,679) of all predicted proteins from the ten analyzed *Aspergillus* strains constituted 13,179 Jaccard orthologous clusters. For *A. nidulans*, 9,250 of its predicted proteins were organized in Jaccard orthologous clusters, meaning that roughly 80% of all *A. nidulans* proteins were linked to at least one ortholog of another *Aspergillus* species. Of the 3,498 GO annotated *A. nidulans* genes, 97% were contained in Jaccard orthologous clusters, meaning that their associated annotations could be mapped to at least one other *Aspergillus* species (see Figure 1). Overall, mapping resulted in an average of 3,484 GO annotated transcripts per genome ranging from 3,403 (*A. clavatus*) to 3,574 (*A. flavus*). On average, their GO annotations comprise 5,436 terms, (see Table 1). These numbers correspond well to the GO annotation of *A. nidulans* and indicate that the majority (97%) of the *A. nidulans* GO annotated genes could be efficiently mapped to the other *Aspergilli*.

Availability of GO resources for the genus *Aspergillus*

The newly mapped GO annotations were deposited at the Broad Institute’s website (http://www.broadinstitute.org/fetgoat/index.html). Different annotation file formats were generated that can be used with diverse public tools for GO enrichment analysis, such as: the Gene Set Enrichment Analysis tool (GSEA) [11], the functional annotation suite Blast2GO [12], the Cytoscape plug-in BiNGO [13] and the Bioconductor package TopGO [14]. To further improve its accessibility, we have implemented Fisher’s exact test [15], a well-accepted approach for GO enrichment analysis, in the web application FetGOat and integrated the newly mapped GO annotations. FetGOat can be accessed via a web
interface at the Broad Institute’s website (http://www.broadinstitute.org/fetgoat/index.html). It combines GO annotations for all Aspergillus species with public genome sequences and a widely used statistical methodology to identify overrepresented GO terms. Via the web interface, a list of gene identifiers can be uploaded to the server and statistical parameters can easily be adjusted with end-user computational skills. After completion of the analysis on the server-side, the enrichment results are sent by Email. The results consist of plain text and spreadsheet files as well as scalable vector graphics representing graphs of enriched GO terms.

Case studies

To demonstrate the flexibility and value of the newly generated resources for omics data analysis, we performed two case studies analyzing transcriptomic datasets recently published for the genus Aspergillus. In the first case study, we demonstrate that the generated resources can be used with various methods for enrichment analysis. We analyze a set of maltose-induced genes from A. niger using FetGOat and two alternative tools for enrichment analysis to subsequently compare their results. In the second case study, we highlight the advantage of having GO annotations that are as comprehensive as possible available for different species. We use FetGOat to analyze sets of glycerol-induced genes derived from a three-species microarray study to highlight major differences in the transcriptional responses for A. nidulans, A. niger and A. oryzae.

Maltose-induced genes

The first dataset reflects the transcriptomic responses of A. niger to growth in maltose and xylose-limited chemostat cultures at identical growth rates. From manual analysis of roughly 700 upregulated genes, Jørgensen et al. [16] concluded a concerted induction of secretory pathway genes in maltose compared to xylose-limited cultures. Using three alternative approaches, we repeated the analysis of the maltose induced genes in an automated and un-biased manner to subsequently compare their enrichment results. First, we performed the analysis using the web application FetGOat. We identified 73 enriched GO terms, which were reduced to 19 most-specific GO terms by removing redundant higher hierarchy terms with less detailed annotations. In correspondence to the findings by Jørgensen et al., the enriched GO terms are related to important steps involved in protein secretion: Translocation to the endoplasmic reticulum, glycosylation and transport between the endoplasmic reticulum and the Golgi apparatus (see Table 2).

For comparison of FetGOat with alternative programs, we used the generated annotation files and repeated the enrichment analysis with two publicly available tools, Blast2GO [12] and GSEA [11]. The numbers of enriched GO terms found with Blast2GO and GSEA are in the same range compared to the results from FetGOat, they identified 76 and 47 enriched GO terms, respectively. To compare the enrichment results from the three tools, we computed semantic similarity scores with the G-SESAME tool [17]. For both FetGOat and Blast2GO, the enrichment statistic is based on Fisher’s exact test and thus their results are theoretically expected to be identical resulting in a semantic similarity score of 1. A similarity score of 0.983 confirms that their results are virtually identical, with minor differences that are likely due to differences in their implementations. In contrast to FetGOat (and Blast2GO), the GSEA results are based on running-sum statistics computed from the complete expression data set. Therefore, the similarity between their results can be expected to be less. Accordingly, G-SESAME determined a smaller semantic similarity score of 0.863 for the results obtained with FetGOat and the GSEA tool.

In addition to the GO terms identified by both Fisher’s exact test based tools, GSEA computed an enrichment of GO terms related to oxidative phosphorylation (GO:0006119), carbohydrate transport (GO:0008643) and glucosidase activity (GO:0015926). Comparing maltose to xylose limitation, an enrichment of those GO
terms fits our expectations. Under maltose-limitation, *A. niger* breaks down the disaccharide into its monomer glucose by enzymes having glucosidase activity. Subsequently, glucose is taken up by carbohydrate transporters, which can be expected to be different from those required for the uptake of xylose. Finally, 1 mole of glucose yields more ATP than 1 mole of xylose, thereby explaining an induction of oxidative phosphorylation.

These differences in the enrichment results are potentially inherited by the statistics applied by Jørgensen *et al.* to define their set of maltose induced genes. In contrast to the GSEA tool, which analyzes the complete expression data, FetGOat and Blast2GO are depending on a-priori performed statistics that were applied to generate subsets of genes or proteins of interest. Jørgensen *et al.* used the Affymetrix MAS 5.0 algorithm for data pre-processing in combination with the student’s t-test to define their set of maltose induced genes. In current literature, this approach is critically discussed [18,19]. To assess the effect of those a-priori applied statistics on the differences between the results from FetGOat and the GSEA tool, we generated an alternative set of maltose-induced genes. We computed RMA expression data [18] from the raw data (CEL files) and subsequently applied a moderated t-statistic [20] to identify upregulated genes (data not shown). Interestingly, FetGOat also identified enriched GO terms related to glucosidase activity and carbohydrate transport for this alternative set of maltose-induced genes. However, no enrichment of genes related to oxidative phosphorylation was found. Genes annotated with the GO term oxidative phosphorylation were only marginally induced and their FDR values were rather high (data not shown). Interestingly, similar differences between Fisher’s exact test based methods and the GSEA tool were reported in another study. In muscle tissue from diabetics, the GSEA tool identified a joint downregulation of genes related to oxidative phosphorylation compared to healthy controls, while no enrichment was found in the set of downregulated genes [21]. For tightly regulated essential cellular processes that show only minor fold changes, the GSEA tool seems to be superior to gene-by-gene differential expression studies.

### Glycerol-induced genes

In the second case study, we used FetGOat to analyze transcriptomic data generated by Salazar *et al.* [22].

| GO term | Description | FDR    | Ontology | Transcripts | Predicted |
|---------|-------------|--------|----------|-------------|-----------|
| **Translocation to ER** | | | | | |
| GO:0031204 | posttranslational protein targeting | 2.7E-04 | BP | 6 | 6 |
| GO:0031207 | Sec62/Sec63 complex | 6.1E-03 | CC | 3 | 3 |
| GO:0005787 | signal peptidase complex | 6.1E-03 | CC | 3 | 3 |
| GO:0006616 | SRP-dependent cotranslational protein targeting | 1.7E-02 | BP | 4 | 5 |
| GO:0051605 | protein maturation by peptide bond cleavage | 1.6E-02 | BP | 5 | 8 |
| **Glycosylation in ER** | | | | | |
| GO:0005788 | endoplasmic reticulum lumen | 3.3E-03 | CC | 4 | 5 |
| GO:0008250 | oligosaccharyltransferase complex | 6.9E-03 | CC | 4 | 6 |
| GO:0006487 | protein amino acid N-linked glycosylation | 4.2E-02 | BP | 8 | 24 |
| GO:0016758 | transferase activity, transferring hexosyl groups | 3.4E-02 | MF | 14 | 54 |
| **Transport between ER and golgi** | | | | | |
| GO:0030126 | COPI vesicle coat | 1.3E-02 | CC | 4 | 7 |
| GO:0030127 | COPII vesicle coat | 6.9E-03 | CC | 4 | 6 |
| GO:0008888 | ER to Golgi vesicle-mediated transport | 4.8E-03 | BP | 22 | 92 |
| GO:0030173 | integral to Golgi membrane | 12.0E-02 | CC | 5 | 12 |
| **Starch metabolism** | | | | | |
| GO:0005982 | starch metabolic process | 4.2E-02 | BP | 5 | 10 |
| **Miscellaneous** | | | | | |
| GO:0006066 | alcohol metabolic process | 4.2E-02 | BP | 33 | 199 |
| GO:0003756 | protein disulfide isomerase activity | 8.2E-03 | MF | 4 | 4 |
| GO:0006083 | acetate metabolic process | 4.2E-02 | BP | 7 | 19 |
| GO:0015812 | gamma-aminobutyric acid transport | 4.2E-02 | BP | 6 | 14 |
| GO:0015935 | small ribosomal subunit | 5.6E-03 | CC | 12 | 44 |

Most specific GO terms identified by FetGOat as being enriched in the maltose-induced gene set. GO terms were grouped in five arbitrary categories: Translocation into endoplasmic reticulum (ER), glycosylation in ER, transport between ER and golgi, starch metabolism and miscellaneous.
With a three-species microarray, the authors studied the transcriptomic responses of *A. nidulans*, *A. niger* and *A. oryzae* to growth in glycerol and glucose-limited batch cultures. The authors identified 4,139 glycerol-induced genes comprising 679, 2,240 and 1,040 genes from *A. nidulans*, *A. niger* and *A. oryzae*, respectively. Based on tri-directional best blast hits, 81 orthologous gene clusters were shown to be upregulated in each of the species. Using the *A. niger* (strain ATCC 1015) GO annotation, Salazar *et al.* analyzed the set of conserved upregulated genes and identified enriched BP terms, which are related to amino acid metabolism, gluconeogenesis, hexose and alcohol biosynthetic processes.

First, we repeated the enrichment analysis similar to Salazar *et al.* on the set of 81 upregulated and conserved genes. With the web application FetGOat, we individually performed enrichment analysis using GO annotations of *A. nidulans*, *A. niger* (strain ATCC 1015) and *A. oryzae*. FetGOat identified 58, 57 and 54 enriched BP terms, respectively. To summarize the enrichment results for the three *Aspergilli* and compare them with each other, we mapped the GO terms to a GO Slim annotation and counted the occurrences of related GO terms. As expected from analyzing orthologous gene sets, the counts for the GO Slim terms were nearly identical, independent of which of the three *Aspergilli* the enrichment analysis was performed for (see Figure 2).

Next, we aimed to identify differences in the tendencies of the transcriptional responses to glycerol for the three *Aspergilli*. With FetGOat, we individually performed enrichment analysis on each of the complete sets of upregulated genes and found 35, 100 and 65 enriched BP terms for *A. nidulans*, *A. niger* and *A. oryzae*, respectively. The differences in the number of enriched BP terms correspond to the differences in the number of upregulated genes. To summarize and compare the results with each other, we mapped the GO terms to a GO-Slim annotation and counted their occurrences (see Figure 3). This summary clearly shows

![Figure 2](image_url)
different tendencies in the transcriptomic responses of the three *Aspergilli*. Most strikingly, a number of GO-Slim terms were identified as being enriched for *A. niger* but not for the other two *Aspergilli*. Many of the associated GO terms are directly or indirectly related to nutrient limitation such as conidiation, secondary metabolic processes and cell death. Furthermore, FetGOat found an enrichment of the BP term response to nutrient levels (GO:0031667) for *A. nidulans* (nine upregulated genes) and *A. niger* (30 upregulated genes) but not *A. oryzae*. In contrast, GO terms related to energy generation and peroxisomal organization were enriched for *A. oryzae* but not for the other two *Aspergilli*. FetGOat further computed an enrichment of the BP term carbohydrate transport (GO:0006843) specifically for *A. oryzae*. Interestingly, the different transcriptional trends correspond well with the physiological data. The capacities to grow on glycerol differ significantly for the three *Aspergilli*. With a maximum specific growth rate of 0.05 h⁻¹, which is one-fourth of its maximum specific growth rate on glucose, *A. niger* grew the worst on glycerol. In contrast, *A. oryzae* showed the fastest growth (0.30 h⁻¹), which is equal to approximately 80% of its maximum specific growth rate on glucose. *A. nidulans* is in between and grew with roughly 50% of its glucose specific speed.

**Discussion**

A detailed and comprehensive genome annotation can be considered a prerequisite for the analysis and interpretation of omics data. GO provides a framework for functional annotation and has been proven to be a valuable tool for omics data analysis, especially in combination with enrichment statistics. Currently, the GO
reference genome project [23] provides the most comprehensive manually curated GO annotation for twelve model organisms and is intended to serve as a reference for automated mapping of GO annotation to organisms other than these major models. From the reference genome projects, Saccharomyces cerevisiae and Schizosaccharomyces pombe are most closely related to the genus Aspergillus.

A. nidulans has so far been the only Aspergillus species with comprehensive genome scale GO annotation based on both orthology mapping to S. cerevisiae and extensive manual curation [9] of gene-specific literature. We have thus mapped the A. nidulans GO annotation to all other Aspergillus species (see Table 1) with published genomes. With 79% of all A. nidulans genes being organized in Jaccard orthologous clusters covering 97% of all its GO annotated genes, we demonstrated that this approach is promising for mapping GO annotation between closely related genomes such as those of the genus Aspergillus. Nevertheless, the newly generated GO annotations have exclusively been inferred by computational analysis and thus their quality can be expected to be lower compared to the extensively manually curated A. nidulans GO annotation. The ortholog clustering approach as implemented in the Sybil comparative analysis package [10] has worked well for a number of comparative genome studies [24-33], but does have limitations, especially when there are a large number of strains and/or percentage of repetitive proteins. Additionally, we recognize that the optimal choice of an ortholog detection method depends on the purpose of the analysis. This graph based approach is robust if looking at closely related species, but may not be the best choice when considering large numbers of more distantly related genomes.

The GO annotations for ten Aspergillus strains (see Table 1) have been made available at the Broad Institute’s website (http://www.broadinstitute.org/fetgoat/index.html) and will be updated regularly as the GO annotations for the various Aspergillus species continue to improve through manual and computational efforts. To improve the applicability of the GO annotations, they are provided in different file formats that can be used with various freely available GO enrichment tools, e.g. Blast2GO [12], TopGO [14], GSEA [11] and BinGO [13]. Thereby, functional analysis of Aspergillus omics data by GO enrichment analysis is strongly facilitated. The availability of different annotation file formats makes it feasible to use different tools and compare them with each other.

To further improve the accessibility of the extended annotations, we developed the web application FetGOat and integrated the GO annotation for all Aspergillus species with public genome sequences. FetGOat basically resembles the functionality of other publicly available enrichment tools. However, for the Aspergillus research community, FetGOat is a valuable addition to existing programs because it uniquely combines an intuitive web interface, GO annotations for all Aspergilli with public genome sequences and a frequently applied statistical method for the identification of enriched GO terms.

To demonstrate the use of those newly generated resources for functional analysis of omics data, we applied them in two case studies to re-analyze recently published microarray data in an automated and unbiased manner. As shown for the first dataset, the enrichment results are in correspondence to the main conclusions from Jørgensen et al. [16]. We found an induction of processes related to secretion, glycosylation and starch degradation (see Table 2). In addition, we used the dataset from Jørgensen et al. to compare the enrichment results of FetGOat to those obtained with two well established publicly available tools, Blast2GO and GSEA. The three tools apply two different methods for enrichment analysis. While Blast2GO and FetGOat compute a Fisher’s exact test statistic to identify GO terms that are over-represented in subsets of genes derived e.g. from transcriptomic or proteomic data, the GSEA tool computes running sum statistics on (non-filtered) expression data to identify a-priori defined groups of genes that show joined differential expression. The results from FetGOat are virtually identical to the results obtained with Blast2GO demonstrating the correctness of FetGOat. As expected, the similarity between the results from FetGOat and the GSEA tool is less, while their results are still well comparable. For a large part, both tools are highlighting the same transcriptional trends. However, the GO term oxidative phosphorylation was exclusively identified as being enriched by the GSEA tool. Taking into account that 1 mole of glucose yields more ATP than 1 mole of xylose, an induction of the oxidative phosphorylation machinery during growth in maltose-limited cultures can be expected. Because the fold-changes of the corresponding genes were very small and their statistical significances were low, no enrichment could be found in the set of maltose-induced genes as assessed by Fisher’s exact test. Similar results were found in another study, in which the GSEA tool detected a joined transcriptional downregulation of genes related to oxidative phosphorylation in tissue from diabetics vs. control [21]. For tightly regulated essential genes, which show only marginal differential expression, the GSEA tool seems to be superior to gene-by-gene differential expression approaches. However, we would like to emphasize that this is rather caused by the a-priori performed statistics than by the Fisher’s exact test itself. A combination of clustering
based on gene expression profiles combined with Fisher’s exact test enrichment statistics will potentially allow to draw similar conclusions as with the GSEA tool. The causality between an increased ATP yield for maltose and an upregulation of secretion related genes remains to be investigated. However, it is an interesting new hypothesis for further investigations.

For the second dataset from Salazar et al. [22], we first performed GO enrichment analysis on the set of 81 conserved and glycerol-induced genes used in the original study. We could partly reproduce the enrichment results. However, we didn’t find an enrichment of genes annotated with the GO term gluconeogenesis. A comparison of the GO annotation used by Salazar et al. and our newly mapped GO annotation revealed that this is due to an improvement of the newly mapped GO annotation, which includes many more genes annotated with the GO term gluconeogenesis. As expected from analyzing orthologous gene sets, we showed that the enrichment results are nearly identical, independent of which of the three Aspergilli they were obtained for. Furthermore, we separately performed enrichment analysis for the three Aspergilli analyzing their complete sets of up regulated genes and highlighted major differences in their responses to glycerol vs. glucose limitation. Thereby, we were able to draw additional conclusions explaining their different capabilities to grow on glycerol. Especially for the three-species microarray platform, FetGOat in combination with the newly mapped GO annotation forms a new, valuable and flexible form, FetGOat in combination with the newly mapped GO annotation to the three species related species to subsequently facilitate the interpretation results. With the different GO annotation files generated in this study, various freely available tools can easily be used and compared with each other. Especially for the genus Aspergillus, FetGOat stands out with respect to the ease of use and the integration of comprehensive and regularly updated GO annotations. The power of FetGOat lies in its flexibility. Any set of genes/proteins from any Aspergillus strain with published genome sequence can be investigated for enrichment of GO terms. FetGOat is not restricted to the genus Aspergillus as it can be extended to include GO annotations from any organism of interest.

Conclusions

We have mapped the A. nidulans GO annotation to the genomes of seven other Aspergillus species and made the GO annotations available in different file formats. We furthermore developed the web tool FetGOat, which can be used for GO enrichment analysis of omics data from all Aspergillus strains with published genome sequences. Both, the mapped GO annotations and FetGOat were successfully applied in two case studies and are available at the Broad Institute’s website (http://www.broadinstitute.org/fetgoat/index.html). Moreover, we have given a general example of how a well annotated genome can help improving GO annotation of related species to subsequently facilitate the interpretation of omics data.

Methods

Ortholog and paralog identification

Clusters of orthologous proteins from ten Aspergillus strains (see Table 1) were generated with Sybil [10]. The Sybil comparative analysis package currently utilizes the following two-step clustering method, which is a modification of the standard reciprocal best match approach. First, an all-vs-all protein similarity matrix is computed by searching each of the predicted polypeptides within the genomes being compared against all polypeptides. BLASTP is currently used for these searches, with an E-value cut off of 1E-5. Polypeptides from each individual species are clustered independently using only BLASTP hits that had a sequence identity score of at least 80%. The BLASTP matches that meet these criteria are used to compute a Jaccard similarity coefficient [34] for each distinct pair of polypeptides in the same genome. Given two polypeptides P1 and P2 the Jaccard similarity coefficient is defined as:

\[ J(P1, P2) = \frac{\text{matches to } P1 \cap P2}{\text{matches to } P1 \cup P2} \]

Using default parameters, any pair of polypeptides with a Jaccard coefficient > 0.6 is connected in a graph.
representation. The connected components of this graph are referred to as “Jaccard Clusters” and are analogous to paralogous protein clusters within each species. Subsequently, the reciprocal best-hit phase of the clustering algorithm identifies pairs of Jaccard clusters such that: (1) The clusters are from different genomes. (2) The highest-scoring BLASTP match of at least one polypeptide in each of the clusters is to a polypeptide in the other cluster. A graph is constructed, with an edge drawn between two nodes (Jaccard clusters) if and only if they are bidirectional best BLASTP matches of each other. The connected components of this graph are considered ortholog groups in downstream analysis and will be referred to as “Jaccard orthologous clusters”.

**Mapping A. nidulans GO annotation**

GO annotation for *A. nidulans* (gene_association.aspgd version: 1.256) was obtained from the Aspergillus Genome Database (AspGD: http://www.aspgd.org) [9] and is based on orthology mapping between *A. nidulans* and *S. cerevisiae* as well as extensive manual curation based on gene specific *A. nidulans* literature. GO terms for Jaccard orthologous clusters and their associated proteins were inferred from *A. nidulans* GO annotation such that each protein belonging to the same Jaccard orthologous cluster shares identical GO terms. For each of the analyzed strains (see Table 1), individual GO annotation files were generated in different formats.

**Enrichment analysis**

GO enrichment analyses were performed applying two different statistical tests: Fisher’s exact test [15] and Kolmogorov-Smirnov statistics [11,35]. If not stated differently, p-values were corrected according to Benjamini & Hochberg [36] and a critical False Discovery Rate (FDR) q-value of 0.05 was applied. For the Fisher’s exact test based enrichment analysis of GO terms, we developed the web application FetGOat, which calculates one-tailed p-values and corrects them for multiple hypothesis testing according to the Benjamini & Hochberg method. In addition to FetGOat, Blast2GO [12] was used to compute enriched GO terms via Fisher’s exact test as implemented in GOSSIP [37]. For the identification of enriched GO terms based on the Kolmogorov-Smirnov statistic, the GSEA tool [11] was used. The corresponding GO annotation files for Blast2GO and the GSEA tool were generated in this study and are available at the Broad Institute’s website (http://www.broad-institute.org/fetgoat/index.html).

**Mapping to GO slim annotation**

To summarize GO enrichment results, we mapped the enriched GO terms to a GO Slim annotation [1], which is a reduced version of the complete annotation with less detailed high-level GO terms, and counted the occurrences (single occurrence option) of GO Slim terms as well as related lower hierarchy terms using the CateGORizer tool [38].

**Acknowledgements**

This work was supported by grants of the SenterNovem IOP Genomics project (IGE07008) and the National Institute of Allergy and Infectious Diseases at the US National Institutes of Health (R01 AI077599). Part of this work was carried out within the research programme of the Kluyver Centre for Genomics of Industrial Fermentation, which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research.

**Authors’ contributions**

BVM implemented a data analysis pipeline to map GO annotation to Jaccard clusters and generate annotation files, developed FetGOat and performed further data analysis. JC and JRW generated Jaccard clusters. GCC integrated FetGOat at the Borad Institute’s website. JRW, AFJR and VM were involved in writing the manuscript. All authors read and approved the final manuscript.

**Received: 4 July 2011 Accepted: 5 October 2011 Published: 5 October 2011**

**References**

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarides A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000, 25:25-9.
2. Jorgensen TR, Nitsche BM, Lamers GE, Aentenshurt M, van den Handel CA, Ram AF. Transcriptomic insights into the physiology of Aspergillus niger approaching a specific growth rate of zero. Appl Environ Microbiol 2010, 76(16):5344-55.
3. Lin M, Lee YJ, Lough TJ, Phinney BS, Lucas W. Analysis of the pumpkin phloem proteome provides insights into angiosperm sieve tube function. Mol Cell Proteomics 2009, 8(2):343-56.
4. Twine NA, Janitz K, Wilkins MR, Janitz M. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer’s disease. PLoS ONE 2011, 6:e18266.
5. Ryan JC, Morey JS, Bottein MY, Rasmussel JS, van Dolah FM. Gene expression profiling in brain of mice exposed to the marine neurotoxin ciguatoxin reveals an acute anti-inflammatory, neuroprotective response. BMC Neurosci 2010, 11:107.
6. Jones MG. The first filamentous fungal genome sequences: Aspergillus leads the way for essential everyday resources or dusty museum specimens? Microbiology 2007, 153:1-9.
7. Wortmann JR, Gilissen JM, Joardar V, Deegan J, Clutterbuck J, Andersen MR, Archer D, Bencina M, Braus G, Coutinho P, von Dohren H, Doonan J, Driessen AJ, Deurek P, Espeso E, Fekete E, Filippi M, Estrada CG, Geyens S, Goldenman G, de Groot PW, Hansen K, Harris SD, Heinekamp T, Helmisdael K, Hennissat B, Hofmann G, Homan T, Hoino T, Horuchi H, James S, Jones M, Karaffa L, Karanu Z, Kato M, Keller N, Kelly DE, Kiel JA, Kim JM, van der Klei L, Kim KS, Kovalchuk A, Krasevec N, Kuboek C, Liu B, Macabba A, Meyer V, Mirabito P, Meski M, Mos M, Mullins J, Nelson DR, Nielsen J, Oakley BR, Osmani SA, Paukla T, Paszewski A, Paulsen J, Pinsky S, Poci P, Purit P, Ram AF, Renn Q, Robelket X, Robson G, Seiboth B, van Solingen P, Specht T, Sun J, Tahan-Talebi N, Takehita N, Ussery D, van Kuyk PA, Visser H, van der Veenvoort PJ, de Vries RP, Walton J, Xiang X, Xiong Y, Zeng AP, Brandt BW, Cornell MJ, van den Holend CA, Visser J, Oliver SG.
15. Fisher R:

14. Adrian Alexa JR:

18. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP:

10. Crabtree J, Angiuoli SV, Wortman JR, White OR:

11. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA,

13. Maere S, Heymans K, Kuiper M:

BiNGO: a Cytoscape plugin to assess

16. Jorgensen TR, Goosen T, Hondel CA, Ram AF, Iversen JJ:

102(43)

enrichment analysis: a knowledge-based approach for interpreting

19. Bourgon R, Gentleman R, Huber W:

20. Smyth GK:

Independent filtering increases
differential expression in microarray experiments.

121(16)

Calculation of P.

21. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J,

Hirschhorn JN, Altshuler D, Groop LC:

Trypanosomatid parasites.

44.

Aspergillus research community.

Sherlock G:

Orvis J, Shah P, Skrzypek MS, Binkley G, Miyasato SR, Wortman JR,

Basturkmen M, Spevak CC, Clutterbuck J, Kapitonov V, Jurka J,

604(5):2 from Contingency Tables, and the

χ² test for independence in

two-way contingency tables. In:

102(4):

The genome

of Babesia bovis and comparative analysis of apicomplexan hemoproteozoa.

PLoS Pathog 2007, 3(3):e1000046.

Brayton KA, Lau AO, Herndon DR, Hannick L, Kappmeyer LS, Berens SJ,

Paulsen I, White O, Ralph SA, Ren Q, Sargeant TJ, Salzberg SL, Stoeckert CJ,

Science 2005, 309(5733):404-9.

El-Sayed NM, Pyli, Björnberg S, Eklund H, Enqvist S, Ahlgren L, Grundström S,

Helander M, Skoog A, Davidsson M, Vainio T, Järvinen H, Anttila H, Smed S,

Smyth G:

Linear models and empirical bayes methods for assessing
differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004, 3:Article 3.

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sthapit V, Lehar J,

Puigserver P, Carlson E, Riddlestrøm M, Lauridsen E, Hous自在, Dali Mj,

Patternson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Landr E, Sottrash JN, Altrhalte H, Droup LC.

PGC-1alpha-responsive genes

are coordinately downregulated in PGCs of mice lackingPGC-1alpha.

Proc Natl Acad Sci USA 2005, 102(43):15455-50.

23. of the Gene Ontology Consortium TRG:

The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. PLoS Comput Biol 2009, 5(7):e1000531.

24. Carlson JM, Ahrén J, Palsson DB, Sullivan SA, Yamamoto MM, Hoffman SL, Wortman JR, Gardner MJ,

Galski MR, Barnwell JW, Fraser-Liggett CM. Comparative genomics of the

neglected human malaria parasite Plasmodium vivax. Nature 2008, 455(7214):757-63.

25. McDonagh A, Fedorova ND, Crabtree J, Yu Y, Kim S, Chen D, Los O, Cairns T, Goldman G, Armstrong-James D, Haynes K, Haas H, Schieff M, May G, Nierman WC, Bignell G. Sub-telomere directed gene expression during initiation of invasive aspergillosis. PLoS Pathog 2008, 4(10):

1000154.

26. Fedorova ND, Khalid N, Joardar VS, Mati R, Amedeo P, Anderson MJ, Crabtree J, Silva JC, Badger JH, Albarnat A, Angiolii S, Bussey H, Bovyer P, Cotty PJ, Dyer PS, Egan A, Galeries K, Fraser-Liggett CM, Haas BI, Imane JM, Kent R, Lemieux S, Malavazi I, Orvis J, Roemer T, Ronning CM, Sundaram JP, Sutton G, Turner G, Venter JC, White OR, Whytl BR, Youngman P, Wolfe KH, Goldman GH, Wortman JR, Jiang B, Denning DW, Nierman WC. Genomic islands in the pathogenic filamentous fungus Aspergillus fumigatus. PLoS Genet 2008, 4(4):e1000464.

27. Brayton K, Lau AO, Herndon DR, Hannick L, Kappmeyer LS, Berens SJ, Bidwield SL, Brown WC, Crabtree J, Fadrosh D, Feldblum T, Forberger HA, Haas BI, Howell JM, Khouri H, Koo H, Mann DJ, Nornine P, Paulsen IT, Radune D, Ren Q, Smith JMK, Suarez CE, White O, Wortman JR, Knowles JDP, McElvain TF, Nene VM. Genome sequence of Babesia bovis and comparative analysis of apicomplexan hemoproteozoa. PLoS Pathog 2007, 3(10):1401-13.

28. Ghedin E, Wang S, Spira D, Caler E, Zhao Q, Crabtree J, Allen JE, Delcher AL, Guliano DB, Miranda-Saavedra D, Angiolii SV, Creasy T, Amedeo P, Haas B, El-Sayed NM, Wortman J, Feldblum T, Tallon L, Schatz M, Shumway M, Koo H, Salzberg SL, Schobel S, Perte P, Mann M, White O, Barton GJ, Carlow CK, Crawford MF, Daub J, Dimmke MW, Estes CF, Foster JM, Granata M, Gregory WP, Kum, Jin J, Kunimura I, KofI, Kumar S, Laney S, Li BW, Li W, Lindblom TH, Lustgarten S, Ma D, Maina CV, Martin DM, McCarbon JP, McElvains L, Mitrava M, Nutman TB, Parkinson J, Peregrin-Alvarez JM, Poole C, Ren Q, Saunders L, Sluder AE, Smith K, Stankie M, Umnash TR, Ware J, Wei AD, Weil G, Williams DJ, Zhang Y, Williams SA, Fraser-Liggett C, Statoia B, Blaxter ML, Scott AA. Draft genome of the filarial nematode parasite Brugia malayi. Science 2007, 317(5845):1766-70.

29. Dunning Hotopp JC, Lin M, Madupu R, Crabtree J, Angiolii SV, Eisen JA, Sehrida R, Ren Q, Wu M, Utterback TR, Smith S, Lewis M, Khouri H, Zhang C, NiU H, Lin Q, Ohsahi N, Zhi N, Nelson W, Brinkack LM, Dodyson RJ, Rosovitz MJ, Sundaram J, Daughey SC, Davidson T, Dunson AK, Gwmino M, Holf DI, Selengut JD, Sullivan SA, Zafar N, Zhou L, Benfielded M, Forberger H, Hafsteinn R, Mulligan S, Robinson J, White O, Rikisha Y, Tettelin H. Comparative genomics of emerging human leishmaniosis agents. PLoS Genet 2006, 2(6):e21.

30. El-Sayed NM, Pyli, Björnberg S, Eklund H, Enqvist S, Ahlgren L, Grundström S, Skoog A, Davidsson M, Vainio T, Järvinen H, Anttila H, Smed S, Smyth G, Shankom S, Jeanl J, Sundaram J, Westenberger S, White O, Melville SE, Donelson JE, Andersson B, Stuart KD, Hall N. Comparative genomics of trypanosomatid parasitic protozoa. Science 2005, 309(5373):404-9.

31. El-Sayed NM, Pyli, Björnberg S, Eklund H, Enqvist S, Ahlgren L, Grundström S, Skoog A, Davidsson M, Vainio T, Järvinen H, Anttila H, Smed S, Smyth G, Shankom S, Jeanl J, Sundaram J, Westenberger S, White O, Melville SE, Donelson JE, Andersson B, Stuart KD, Hall N. Comparative genomics of trypanosomatid parasitic protozoa. Science 2005, 309(5373):404-9.
Domingo AR, Wasawo D, Crabtree J, Wortman JR, Haas B, Angiuoli SV, Creasy TH, Lu C, Suh B, Silva JC, Utterback TR, Feldblyum TV, Perna T, Allen J, Nierman WC, Taracha EL, Salzberg SL, White OR, Fitzhugh HA, Morzaria S, Venter JC, Fraser CM, Nene V. Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes. Science 2005, 309(5731):134-7.

33. Joardar V, Lindeberg M, Jackson RW, Selengut J, Dodson R, Brinkac LM, Daugherty SC, Deboy R, Durkin AS, Giglio MG, Madupu R, Nelson WC, Rosovitz MJ, Sullivan S, Crabtree J, Creasy T, Davidsen T, Haft DH, Zafar N, Zhou L, Halpin R, Holley T, Khouri H, Feldblyum T, White O, Fraser CM, Chatterjee AK, Cartinhour S, Schneider DJ, Mansfield J, Collmer A, Buell CR. Whole-genome sequence analysis of Pseudomonas syringae pv. phaseolicola 1448A reveals divergence among pathovars in genes involved in virulence and transposition. J Bacteriol 2005, 187(18):6488-98.

34. Jaccard P. Nouvelles recherches sur la distribution florale. Bulletin de la Societe Vaudense des Sciences Naturelles 1908, 44:223-270.

35. Hollander M, Wolfe D. Nonparametric Statistical Methods. New York:Wiley, 1999.

36. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 1995, 57:289-300.

37. Bluthgen N, Brand K, Cajavec B, Swat M, Hezel H, Beule D. Biological profiling of gene groups utilizing Gene Ontology. Genome Inform 2005, 16:106-15.

38. H ZL, B J, R JM. CateGOrizer: A Web-Based Program to Batch Analyze Gene Ontology Classification Categories. Online Journal of Bioinformatics 2008, 92(1):108-112.

39. Hulsen T, de Vlieg J, Alkema W. BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genomics 2008, 9:488.

40. Mabey JE, Anderson MJ, Giles PF, Miller CJ, Attwood TK, Paton NW, Bornberg-Bauer E, Robson GD, Oliver SG, Denning DW. CADRE: the Central Aspergillus Data Repository. Nucleic Acids Res 2004, 32 Database: D401-5.

doi:10.1186/1471-2164-12-486
Cite this article as: Nitsche et al. New resources for functional analysis of omics data for the genus Aspergillus. BMC Genomics 2011 12:486.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit