Accurate prediction of secondary metabolite gene clusters in filamentous fungi

Mikael R. Andersen,a,1 Jakob B. Nielsen,a, Andreas Klitgaard,a, Lene M. Petersen,a, Mia Zachariassen,a, Tilde J. Hansen,a, Lene H. Blicher,a, Charlotte H. Gottfredsen,a Thomas O. Larsen,a Kristian F. Nielsen,a, and Uffe H. Mortensen,a

Center for Microbial Biotechnology, Department of Systems Biology, bDTU Multi-Assay Core, Department of Systems Biology, and cDepartment of Chemistry, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark

AUTHOR SUMMARY

Secondary (nongrowth associated) metabolites (SMs) are chemical entities found primarily in plants, fungi, and microbes. SMs comprise molecules such as hormones, antibiotics, and toxins and provide abundant sources of pharmaceuticals (1). Here, we describe methods for predicting and identifying the genes of microbial fungi responsible for the abundant biosynthesis of SMs. These methods are capable of accelerating elucidation of the important SM biosynthetic pathways and should benefit development of pharmaceuticals and synthetic biochemistry (2).

Filamentous fungi are particularly interesting as sources of SMs. Despite their relatively small genomes (30–40 Mb), microbial fungi contain more than 40 different genes catalyzing the biosynthesis of SMs. The number of different compounds produced by each fungus can exceed the number of genes by many times. This increased diversity is due to the highly modular mode of the biosynthesis of SMs, which involves different classes of polymer backbones being modified by a plethora of tailoring enzymes, such as (de)hydratases, oxygenases, hydrolases, and methylases.

In the present study, we collected and expanded a compendium of gene expression data for a model fungus, Aspergillus nidulans, to encompass >40 samples from a diverse set of conditions. We combined the expression profiles with the chromosomal location of the genes to identify colocalized and coregulated genes. Using a statistical method, we identified the member genes of biosynthetic clusters around predicted and known SM synthases. Here, we predicted the members of 58 gene clusters and validated these predictions through comparison with 16 known clusters (see example in Fig. P1). We constructed additional gene deletion strains to investigate further the accuracy of predictions and to compare the findings with the findings of previous studies, as well as to account for changes in gene annotation over time. Our analysis showed overall accuracy of the predictions. The efficiency of the method depends on the number and diversity of the sampling conditions included in gene expression analysis. This diversity should at least include different growth media and liquid as well as solid-state cultivation. The method is immediately applicable to any fungal species with legacy gene expression data and a sequenced genome.

Further, we showed that the gene expression profiles of key genes can be used to predict gene clusters located on different chromosomes involved in the biosynthesis of the same class of compounds (cross-chemistry). Our analysis showed a high degree of coordinated expression between biosynthetic gene clusters, which, in some cases, suggests cross-chemistry between clusters. For example, we used gene deletions and chemical analysis of deletion mutants efficiently to determine two gene clusters on separate chromosomes involved in producing the same family of compounds. We further confirmed the interaction of these gene clusters by structural elucidation of the main compound, a prenylated nonribosomal cyclopeptide called nidulanin A.

In summary, our present findings can immediately support an area of intense focus within fungal biology, namely, the identification of gene clusters involved in biosynthesis of bioactive metabolites, by providing targets and predictions for gene clusters for the important model fungus A. nidulans. Further, in the short term, they provide a general method for rapid prediction of gene clusters in other fungi. This

Fig. P1. Comparison of the gene cluster known to be required for biosynthesis of emericellamide to the predictions of the described method. (A) Gene expression plots of the five genes, AN2545–AN2549, known to be required for emericellamide biosynthesis. (B) Chromosomal map of the emericellamide gene cluster and surrounding genes. The clustering score (CS) evaluating co-regulation is shown for the genes in the columns and in the numbers above the columns. Note how the expression pattern of AN2546, which does not contribute to emericellamide biosynthesis (3), deviates with a statistically insignificant clustering score (CS < 2.13). Genes surrounding the cluster exhibit dissimilar expression patterns (expression values not shown). NRPS, nonribosomal peptide synthetase; PKS, polyketide synthase.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The gene expression data, gene expression microarray data description, and legacy gene expression data reported in this paper are available from the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession nos. GSE39993, GPL15899, GSE12859 and GSE7295).

1To whom correspondence should be addressed. E-mail: mra@bio.dtu.dk.

See full research article on page 899 of www.pnas.org.

Cite this Author Summary as: PNAS 10.1073/pnas.1205532110.
will assist in the identification of biosynthetic genes for a given SM, which can support pathway elucidation in general and is of particular interest for known and potential bioactive compounds. This method can be applied directly to the many fungal species for which large amounts of legacy data exist.