

## The Aspergillus Mine - publishing bioinformatics.

**Vesth, Tammi Camilla; Rasmussen, Jane Lind Nybo; Theobald, Sebastian; Kjærboelling, Inge; Frisvad, Jens Christian; Nielsen, Kristian Fog; Lyhne, Ellen Kirstine; Kogle, Martin Engelhard; Kuo, Alan; Riley, Robert; de Vries, R.P.; Grigoriev, Igor V.; Mortensen, Uffe Hasbro; Henrissat, Bernard; Baker, Scott E.; Andersen, Mikael Rørdam**

*Link to article, DOI:*

[10.7490/f1000research.1113115.1](https://doi.org/10.7490/f1000research.1113115.1)

*Publication date:*

2016

*Document Version*

Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

*Citation (APA):*

Vesth, T. C., Rasmussen, J. L. N., Theobald, S., Kjærboelling, I., Frisvad, J. C., Nielsen, K. F., ... Andersen, M. R. (2016). The Aspergillus Mine - publishing bioinformatics.. Poster session presented at 15th European Conference on Computational Biology, Den Haag, Netherlands. DOI: 10.7490/f1000research.1113115.1

## DTU Library

Technical Information Center of Denmark

---

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# The Aspergillus Mine - Publishing bioinformatics

Tammi C. Vesth\* (1), Jane L. Nybo (1), Sebastian Theobald (1), Inge Kjærboelling (1), Jens C. Frisvad (1), Kristian F. Nielsen (1), Ellen K. Lyhne (1), Martin E. Kogle (1), Alan Kuo (3), Robert Riley (3), R.P. de Vries (4), Igor V. Grigoriev (3), Uffe H. Mortensen (1), Bernard Henrissat (5), Scott E. Baker (2), Mikael R. Andersen (1)

Curious about the Aspmine? Ask me!



Tammi Vesth  
tcve@bio.dtu.dk

- 1) Department of Systems Biology, Technical University of Denmark, Kgs. Lyngby, Denmark
- 2) Joint Bioenergy Institute, Berkeley, CA, USA
- 3) Joint Genome Institute, Walnut Creek, CA, USA
- 4) Fungal Physiology, CBS-KNAW Fungal Biodiversity Centre, Utrecht, the Netherlands



Technical University of Denmark



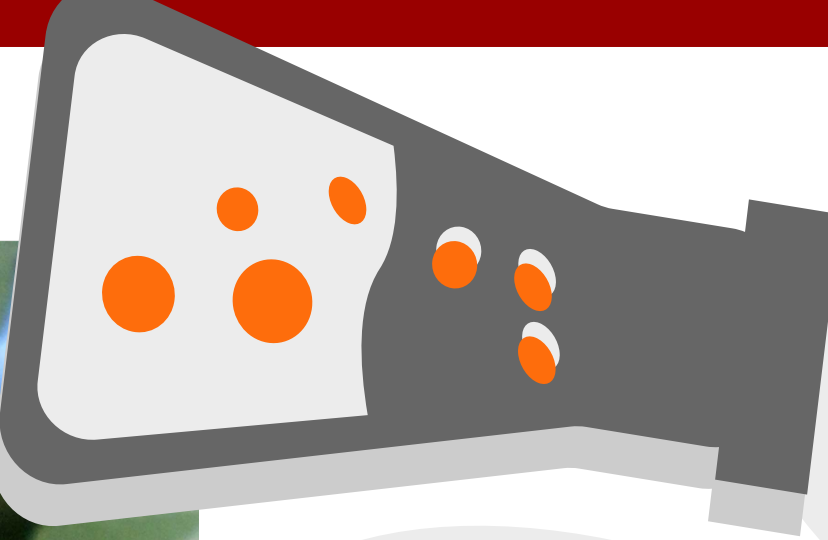
CBS-KNAW Fungal Biodiversity Centre  
An Institute of the Royal Netherlands Academy of Arts and Sciences



Joint BioEnergy Institute



## Background



### Fungi

Producers of **chemically** and **medically** relevant **compounds**

### Aspergilli

Well studied **production organisms**

Can be genetically optimized to produce **cheaper** and environmentally friendly **medicin** and **chemicals**



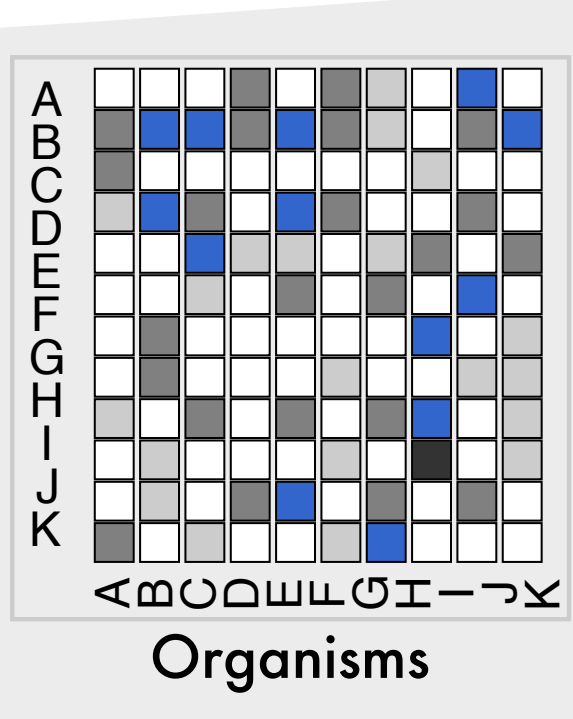
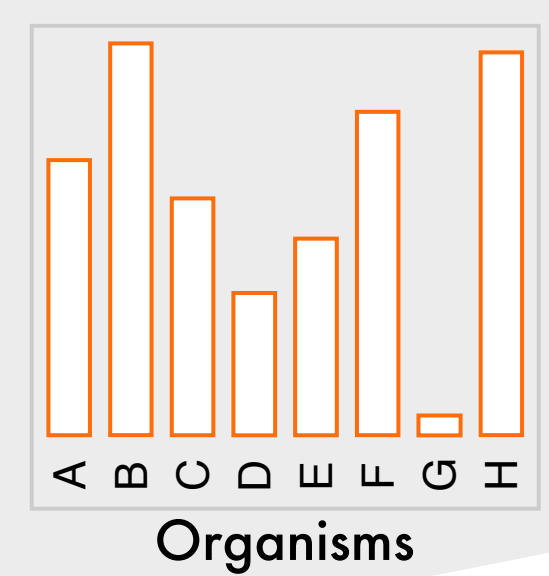
### Sequencing

Initiative to sequence **> 300 species**  
**200 million years of evolution**

**DTU IBT Culture Collection**  
**> 35,000 fungal cultures**  
Penicillium, **Aspergillus**, Fusarium, Alternaria and Trichoderma

### Comparative Genomics

- Shared genes/functions
- Gene content/genome statistics
- Species specific genes
- Secondary metabolism clusters
- Horizontal gene transfers
- Families of genes and clusters
- Sequence data quality
- Genome sequence diversity



### Knowledge sharing

- Cheap hosting of **Interactive** web-applications
- Results** and **analysis** with customization
- Comprehensive **documentation**



**ASPMINE**

- Analysis
- Introduction
- Contact
- Collaborators

**Genomes and proteomes**

**Secondary Metabolism**

**Genetic Diversity**

**GENETIC DIVERSITY**

Here we present a method for comparing shared protein families across a large set of Aspergillus species.

**ANALYZE PAN-CORE GENOMES IN ASPERGILLI**

The Aspergillus is one of the most diverse and important industrial and agricultural filamentous fungus. It has been found in various environments including extreme conditions, and shows a surprising capacity for environmental adaptation while producing a variety of secondary metabolites. The phenotypic diversity seen in species of Aspergillus must, as all other biological diversity, be the product of evolution, including positive selection, horizontal gene transfers and gene rearrangement. Here we discuss the genomic diversity among 37 species by investigating three different patterns in genes.

1. The pan-genome comprises all gene present in one or more species and are expected to code for genetic abilities such as environmental adaptation factors and growth abilities, assuming these are not

## Documentation

- Data science in biology is a increasing
  - Publication of thorough data methods is insufficient
  - Difficulties in reproduction of results
  - The Aspergillus Mine offers online documentation for each analysis
  - Online access to analysis data and descriptions of methods
- aspmine.wordpress.com

## Genetic Diversity

- DNA/protein sequence comparisons are essential to comparative genomics
  - Similar sequences imply similar biological function
  - Proteins with similar sequences form clusters of functionally related proteins
  - Core: Families with one or more proteins from each strain
  - Unique: Families with proteins from only one strain
  - Closely related strains share more families. Many families are strain specific!
- fungiminions.shinyapps.io/geneticDiversity

## Families of proteins

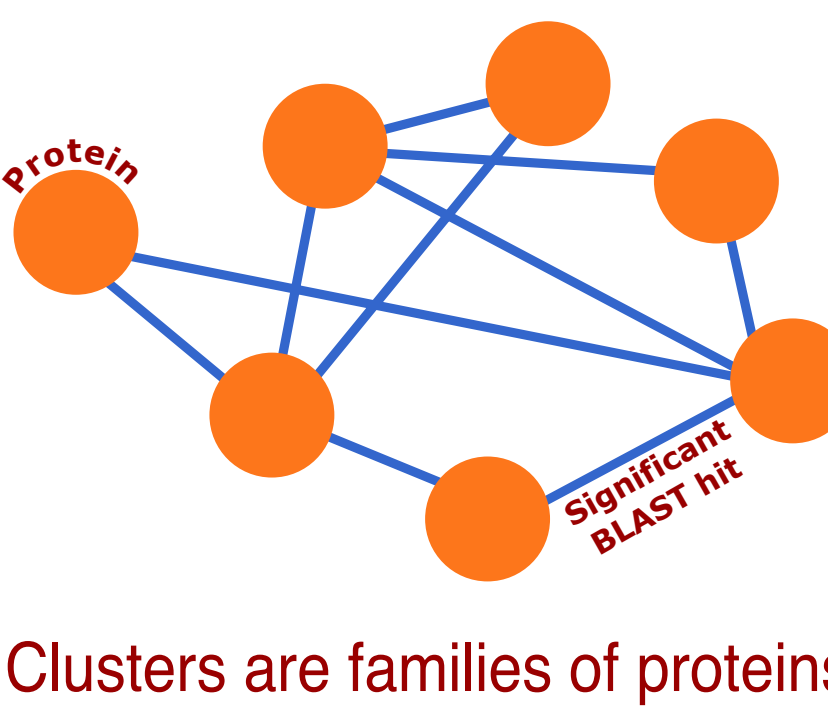
Sequence alignments - BLAST

Q-K-E-S-G-P-S-S-S-Y-C  
V-Q-Q-E-S-G-L-V-R-T-T-C

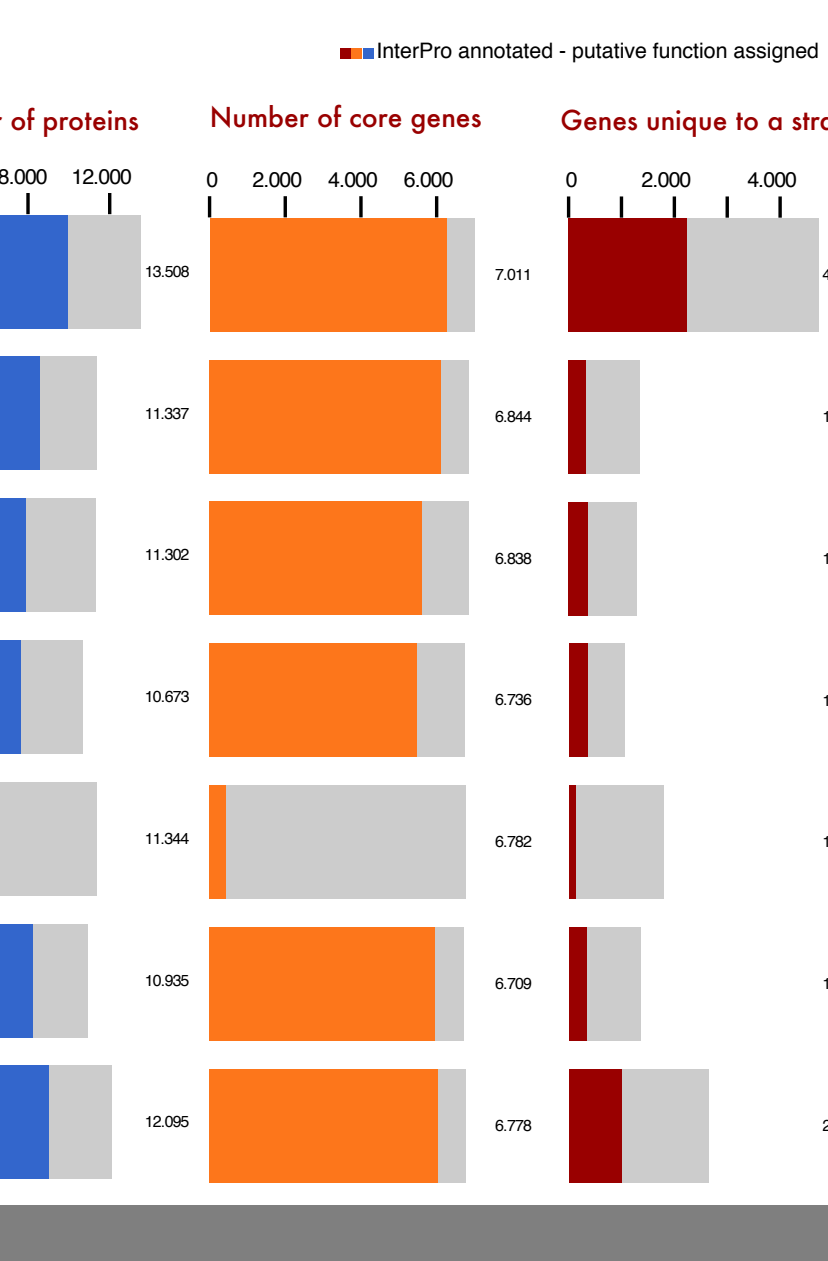
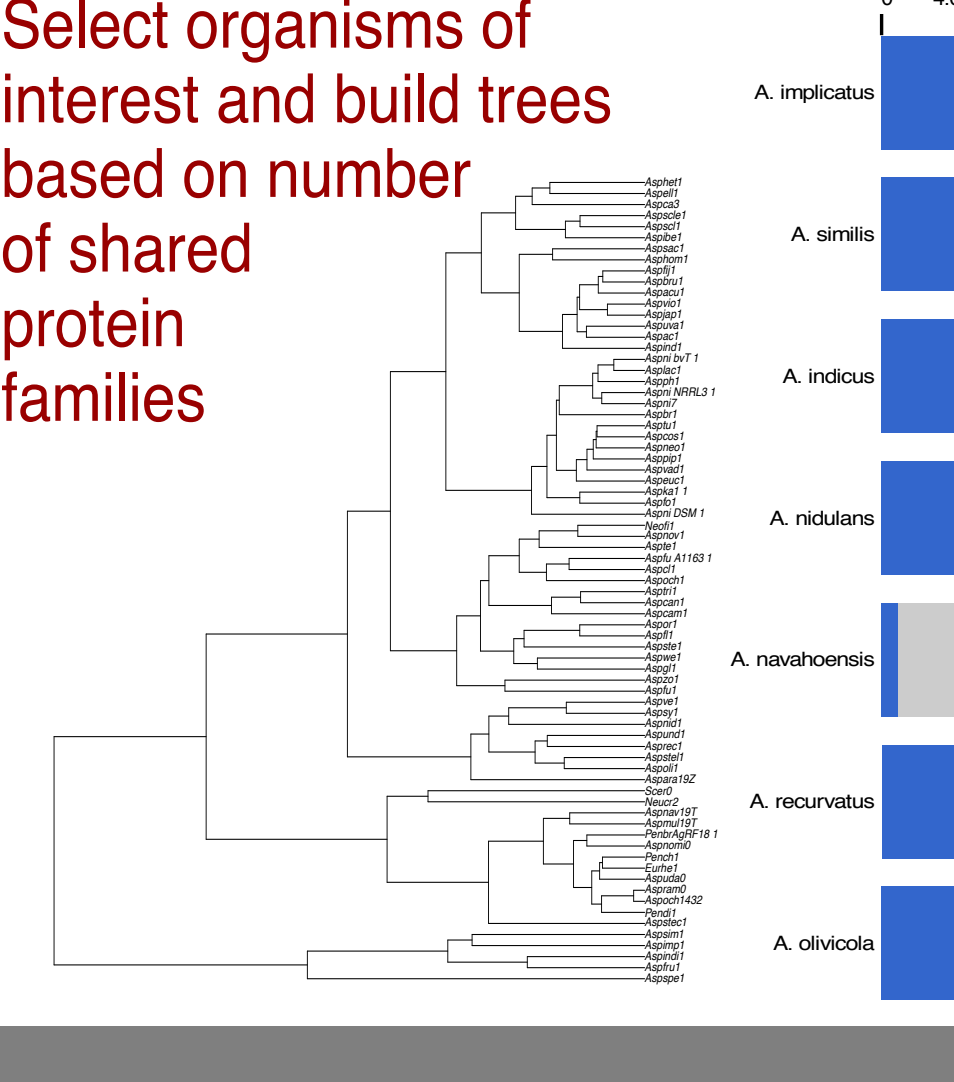
Alignment significance cutoffs

Coverage hit + query > 130% & Percent identity > 50%

## All against all protein alignments and single linkage clustering



## Family conservation



## Approach - connecting data

### Relational database

Strain information

- Raw sequences
- Taxonomy

Functional tables & annotation

- Gene Ontology
- KEGG pathways
- KOG groups
- InterPro domains
- PFAM domains
- SMURF secondary metabolism

BLAST alignment scores

- All against all alignments
- alignment coverage > 50%
- Query/hit coverage
- Percent identity
- Reciprocal hit flag

### Analysis

- Conditional data selection
- Data joining
- Comparative measures
- Protein families
- Cluster families

Specific bindingsite

Specific cellular localization

Homologs in set of species

Specific function

Found in species set

Proximity to other gene

PLATFORM: MySQL, Python & R

PLATFORM: R & RShiny

### Visuals & Publication

Documentation & analysis

Free online hosting with Wordpress

### Data download, figures & interactive figures:

Rshiny hosting with ShinyIO, \$440/Year

- Unlimited Applications
- 500 Active Hours
- Performance Boost

### RShiny web app analysis setup, visuals & downloads

Select organisms and enzymatic activity and view matching clusters. Synteny view of members of the cluster family

## Secondary Metabolism

- Secondary metabolites are important medical and chemical compounds
  - Encoded by gene clusters
  - Clusters are grouped into families
  - Gene clusters are highly diverse, do not follow taxonomy and illustrate mechanisms for speciations
- fungiminions.shinyapps.io/familiesPhyloTreeApp

