Genetic diversity of 100+ Aspergillus species: The aspMine analysis resource

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

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):

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.
Aspergillus species are highly diverse and represent large evolutionary distance.

Species and diversity of secondary metabolism

Comparative genomics can be used to investigate a large number of Aspergillus species. Here we aim to sequence a large number of aspects of genetic diversity. Comparative genomics can be used to investigate a large number of aspects of genetic diversity. Here we aim to sequence a large number of aspects of genetic diversity.

Sequencing

- Families of proteins with shared functions
- Phylogenetic distances
- Genes/proteins specific to single species
- Genetic diversity of Aspergilli
- Shared protein families within groups of species
- Large natural diversity
- Producers of chemically well-studied compounds
- Genetic diversity of 100+ Aspergillus species
- Families of related proteins - aspmine hfams
- Genes found only in a subset of species
- Specific functional annotation which is always found twice in Aspergilli
- Genes containing signal peptides and found in a specific set of species

Analysis database - the workhorse

- Metabolites databases are ideal for comparing different data sources as well as conditional selection of data
- Raw data is obtained from JGI and is stored in a costum-designed relational database (MySQL)
- Data include functional annotation and protein sequences
- All proteins are compared across the growing dataset (BLAST)
- BLAST alignment scores are stored in the database

- MySQL is ideal for extracting data which fulfills specific criteria
- Examples of conditional selection of data in the analysis database
- Genes found only in a subset of species
- Specific functional annotation which is always found twice in Aspergilli
- Genes containing signal peptides and found in a specific set of species

Analysis setup

- Raw information
  - Taxonomical section
  - Protein sequences
  - Organism name and ID
- Functional annotation
  - SMURF - secondary metabolism
  - InterPro - domains
  - KOG - categories
  - KEGG - pathways
- BLAST alignment data
  - Query/hit coverage
  - Alignment coverage
  - Percent identity
  - Reciprocal hit flag

Selecting data conditionally

- Collecting all the data from which we want to query
- MySQL is ideal for extracting data which fulfills specific criteria
- Examples of conditional selection of data in the analysis database
- Genes found only in a subset of species
- Specific functional annotation which is always found twice in Aspergilli
- Genes containing signal peptides and found in a specific set of species

Families of related proteins - aspmine hfams

- Assumption: similar protein sequences imply similar function
- Sequence similarity can be defined by alignments
- Alignments are determined using BLAST
- Alignment coverage of hit plus query must be over 130%
- Identical residues in alignment must be over 50%
- Proteins are connected using single linkage, a protein need only be connected to one other member of the cluster.

Families of related SMGCs - Secondary Metabolism Gene Clusters

- Assumption: SMGCs with similar genes will create similar compounds
- Functionally similar genes can be identified using BLAST
- Enzymes that initiate a secondary metabolite (backbones) are most important in the definition of that metabolite
- Tailoring enzymes also hold information about the metabolite but to a lesser extend than backbone enzymes
- SMGCs must share a significant fraction of backbone and tailoring enzyme activities to create similar compounds
- Pairwise SMGC similarity score
  - Pairwise SMGC similarity score
  - Prior to 80% pairwise SMGC similarity score
  - Prior to 85% pairwise SMGC similarity score
- SMGCs are predicted using the SMURF algorithm
- A costum score is used to calculate the similarity between clusters
- Two rounds of random walk clustering connects the most similar clusters and creates a network which illustrates the interconnectedness of SMGCs

Interactive apps

- The aspMine is a webpage
- Contains documentation of analysis
- Holds link to interactive analysis apps
- Apps allow the user to explore the analysis data
- Figures and tables are available for download
- Analysis in the apps can be customized by selecting organisms of interest and cutoffs or subsets of data

Documented apps

- Analysis apps are available for analysis of genetic diversity through the construction of protein families - aspmine hfams
- DNA/protein sequence comparisons are essential to comparative genomics
- Proteins with similar sequences form clusters of functionally related proteins - protein families
- Closely related strains share more families
- Many families are strain specific

Secondary Metabolism

- Analysis apps are available for analysis of genetic diversity through the construction of protein families and cluster variation across species of Aspergilli
- SMGC families can be queried using protein FASTA sequences or JGI protein identifiers
- Families can be inspected by synteny plots illustrating conserved functions and organization among species
- SMGC clusters do not follow the standard taxonomy and their presence across phylogeny can be explored appropriately
- Hosted with www.shinyapps.io ($440 USD/year)
- Comparison of distances and shared gene cluster families
- Genes unique to a strain

Documentation

- Documentation of data analysis is often neglected
- Publication of thorough data methods is insufficient
- A webpage is a good place to document analysis
- The Asp Mine offers online documentation
- Access to analysis data and descriptions of methods

Fungi

- Producers of chemically well-studied compounds
- Large natural diversity and high production of bioactive compounds
- Well studied production organisms
- Can be genetically optimized for production of cheaper and environmentally friendly compounds

Sequencing

- Genome sequencing and analysis can elucidate many interesting genetic features. In this project we aim to sequence a large number of Aspergillus species
- Initiative to sequence > 300 species
- DTU IBT Culture Collection > 35,000 fungal cultures of Penicillium, Aspergillus, Fusarium, Alternaria and Trichoderma
- Data represents 200 million years of evolution

Comparative Genomics

- Comparative genomics can be used to investigate a number of aspects of genetic diversity. Here we focus on evolutionary development of Aspergillus species and diversity of secondary metabolism
- Genomes/annotation measures and quality
- Genetic diversity of Aspergilli
- Families of proteins with shared functions
- Genes/proteins specific to single species
- Shared protein families within groups of species
- Horizontal gene transfers across large phyllogenetic distances
- Families of secondary metabolism gene clusters responsible for similar compounds

Genetic Diversity

- Analysis apps are available for analysis of genetic diversity through the construction of protein families - aspmine hfams
- DNA/protein sequence comparisons are essential to comparative genomics
- Proteins with similar sequences form clusters of functionally related proteins - protein families
- Closely related strains share more families
- Many families are strain specific

Background

- We live in the digestive tract of the fungi
- Fungi grow in many different environments

The aspMine analysis resource

- https://aspmine.org
- Interactive analysis apps
  - Analysis with customization and documentation
  - Cheap hosting of interactive web-applications
  - Rshiny is a R package for interactive web apps
  - www.shinyapps.io
  - Documentation of data analysis is often neglected
  - Access to analysis data and descriptions of methods

Tamus C. Vestling 1, Jane L. Nybroe 1, Sebastian Thefeldt 1, Inge Karshälling 1, Jens C. Friisad 1, Kristin F. Nielsen 1, Ellen K. Lyhne 1, Martin E. Kogge 1, Adam Kau 2, Robert Riley 2, R.P. de Vries 4, Igor V. Grigoriev 3, Uffe H. Mortensen 1, Bernard Henrissat 5, Scott E. Baker 2, Mikael R. Andersen 1

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, Westerdijk Fungal Biodiversity Institute - KNW Fungal Biodiversity Centre, Utrecht, the Netherlands

Rshiny is a R package for interactive web apps
www.shinyapps.io

Curious about the Aspmine?

Families of protein families - aspmine hfams

- Shared proteins families within groups of species
- Large natural diversity
- Producers of chemically well-studied compounds
- Genetic diversity of 100+ Aspergillus species
- Families of related proteins - aspmine hfams
- Genes found only in a subset of species
- Specific functional annotation which is always found twice in Aspergilli
- Genes containing signal peptides and found in a specific set of species

Genes unique to a strain

- Percent identity
- BLAST alignment data
- Query/hit coverage
- Alignment coverage
- Percent identity
- Reciprocal hit flag

Analysis with customization and documentation

- Hosted with www.shinyapps.io ($440 USD/year)
- Unlimited apps, 500 active hours