Ongoing progress in sequencing fungal genomes and application to fungal population biology and ecology

---

Jason E Stajich
Plant Pathology & Microbiology
University of California, Riverside

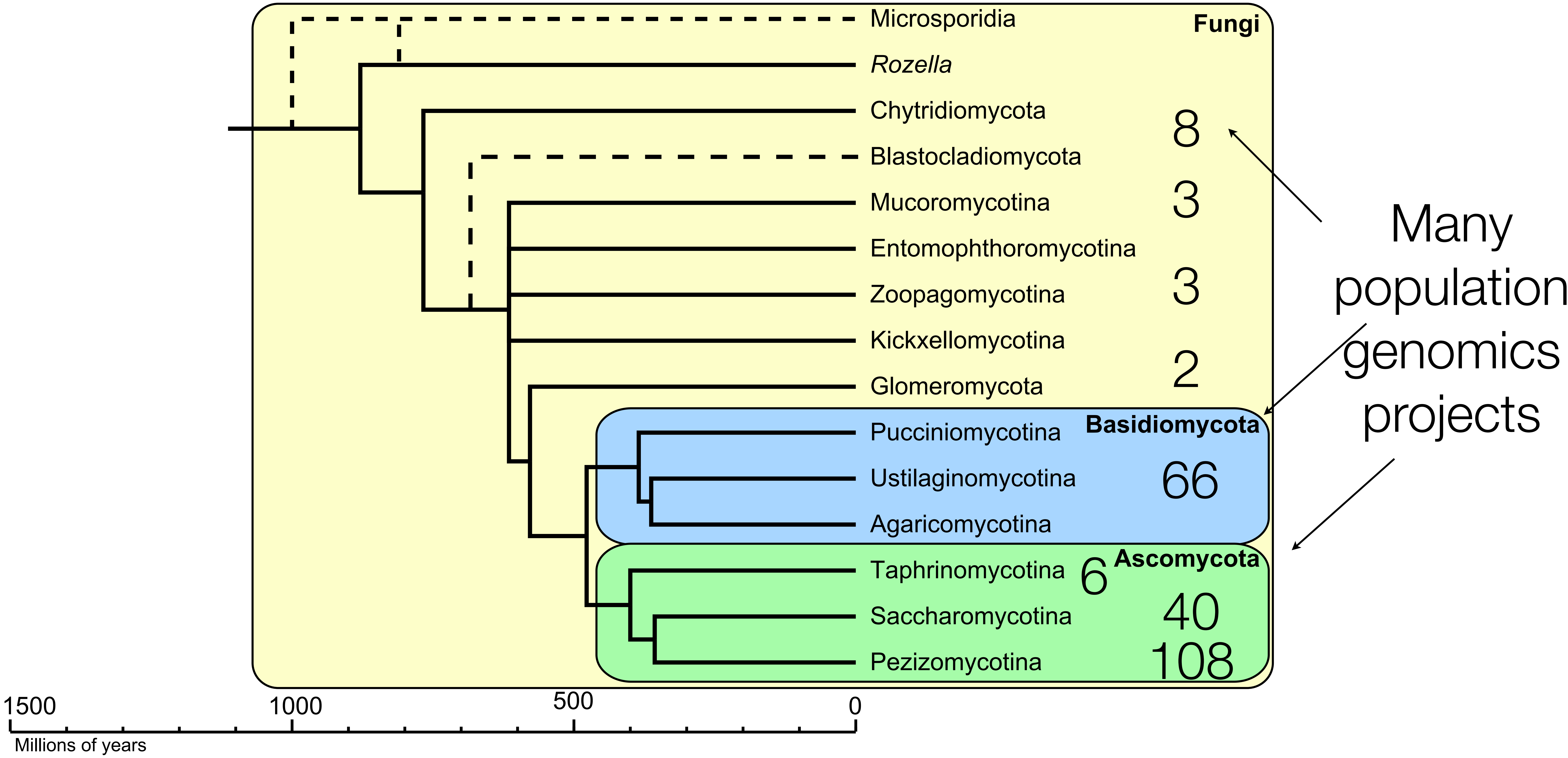
UC RIVERSIDE UNIVERSITY OF CALIFORNIA | Institute for Integrative Genome Biology

CePCeB
Center for Plant Cell Biology
at UC Riverside

Overview

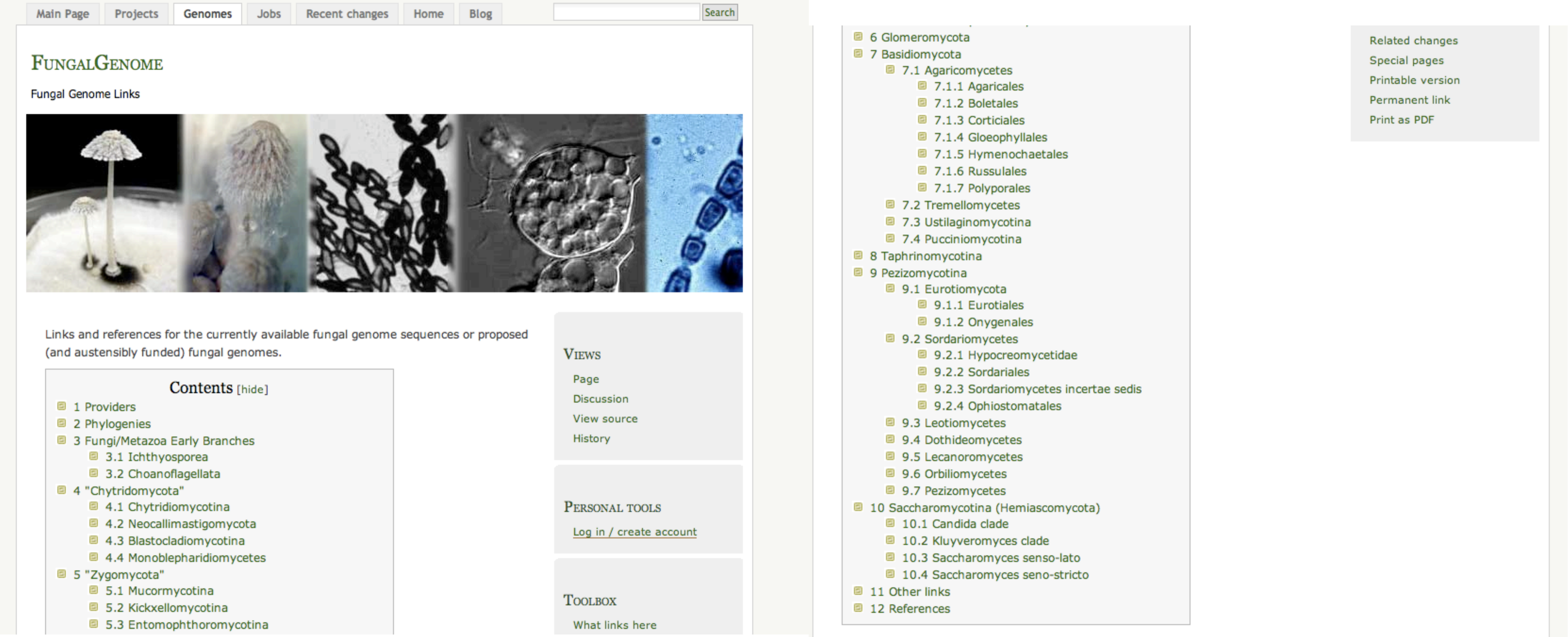
- Fungal Genomes Sequencing progress
- Places to get a genome sequenced for free
- Should you go sequence genomes yourself?
- Ecological and population applications

Genomes from across the kingdom



Many population genomics projects

http://fungalgenomes.org/genomes



FUNGALGENOME

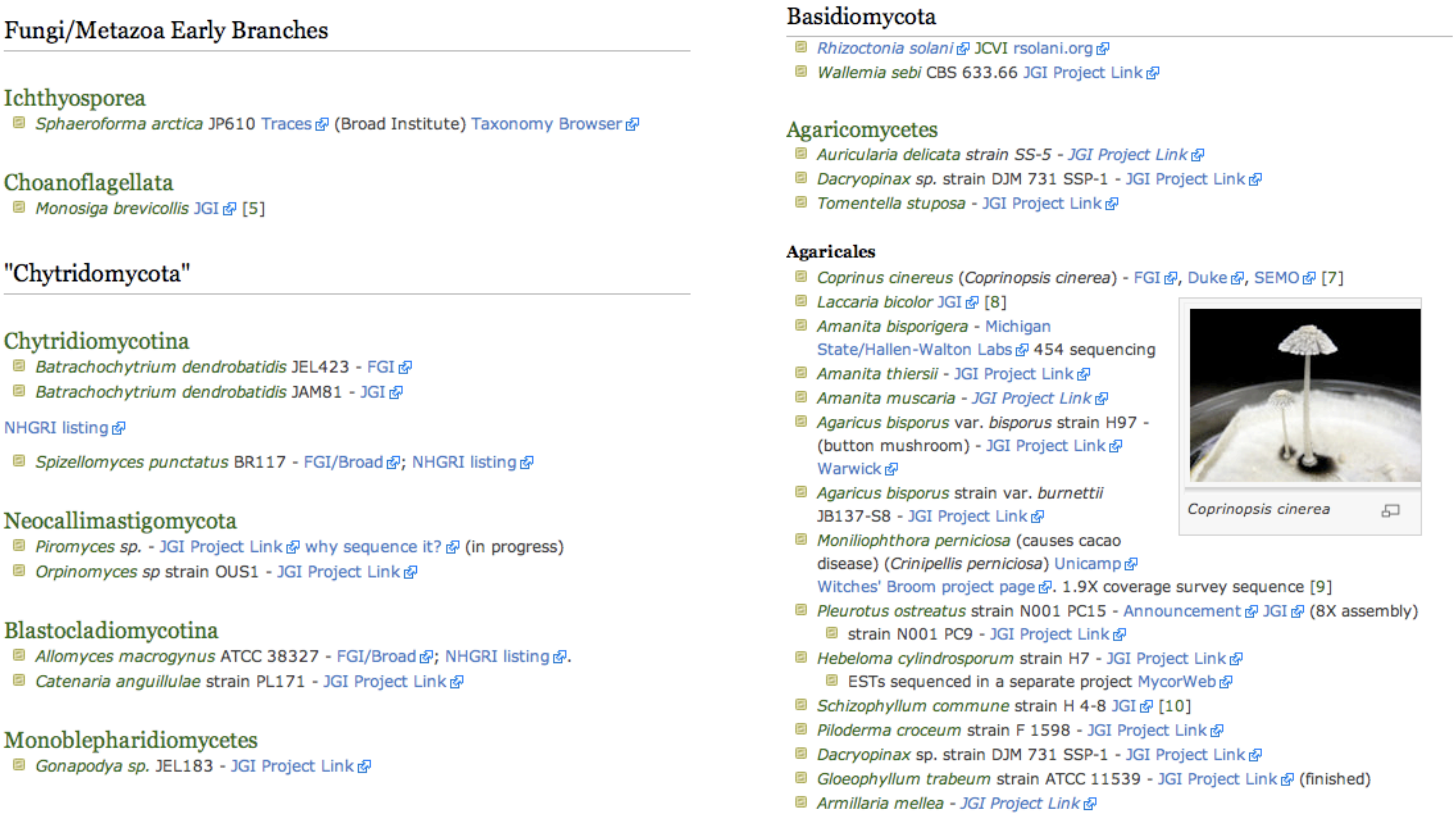
Fungal Genome Links

Links and references for the currently available fungal genome sequences or proposed (and austensibly funded) fungal genomes.

Contents [hide]

- 1 Providers
- 2 Phylogenies
- 3 Fungi/Metazoa Early Branches
  - 3.1 Ichthyosporea
  - 3.2 Choanoflagellata
- 4 "Chytridomycota"
  - 4.1 Chytridiomycotina
  - 4.2 Neocallimastigomycota
  - 4.3 Blastocladiomycotina
  - 4.4 Monoblepharidiomycetes
- 5 "Zygomycota"
  - 5.1 Mucormycotina
  - 5.2 Kickxellomycotina
  - 5.3 Entomophthoromycotina
- 6 Glomeromycota
- 7 Basidiomycota
  - 7.1 Agaricomycetes
    - 7.1.1 Agaricales
    - 7.1.2 Boletales
    - 7.1.3 Corticiales
    - 7.1.4 Gloeophyllales
    - 7.1.5 Hymenochaetales
    - 7.1.6 Russulales
    - 7.1.7 Polyporales
  - 7.2 Tremellomycetes
  - 7.3 Ustilaginomycotina
  - 7.4 Pucciniomycotina
- 8 Taphrinomycotina
- 9 Pezizomycotina
  - 9.1 Eurotiomycota
    - 9.1.1 Eurotiales
    - 9.1.2 Onygenales
  - 9.2 Sordariomycetes
    - 9.2.1 Hypocreomycetidae
    - 9.2.2 Sordariales
    - 9.2.3 Sordariomycetes incertae sedis
    - 9.2.4 Ophiostomatales
  - 9.3 Leotiomycetes
  - 9.4 Dothideomycetes
  - 9.5 Lecanoromycetes
  - 9.6 Orbiliomycetes
  - 9.7 Pezizomycetes
- 10 Saccharomycotina (Hemiascomycota)
  - 10.1 Candida clade
  - 10.2 Kluyveromyces clade
  - 10.3 Saccharomyces senso-lato
  - 10.4 Saccharomyces seno-stricto
- 11 Other links
- 12 References

http://fungalgenomes.org/genomes



Fungi/Metazoa Early Branches

Ichthyosporea

- *Sphaeroforma arctica* JP610 Traces (Broad Institute) Taxonomy Browser

Choanoflagellata

- *Monosiga brevicollis* JGI [5]

"Chytridomycota"

Chytridiomycotina

- *Batrachochytrium dendrobatidis* JEL423 - FGI
- *Batrachochytrium dendrobatidis* JAM81 - JGI

NHGRI listing

- *Spizellomyces punctatus* BR117 - FGI/Broad; NHGRI listing

Neocallimastigomycota

- *Piromyces sp.* - JGI Project Link why sequence it? (in progress)
- *Orpinomyces sp* strain OUS1 - JGI Project Link

Blastocladiomycotina

- *Allomyces macrogynus* ATCC 38327 - FGI/Broad; NHGRI listing.
- *Catenaria anguillulae* strain PL171 - JGI Project Link

Monoblepharidiomycetes

- *Gonapodya sp.* JEL183 - JGI Project Link

Basidiomycota

- *Rhizoctonia solani* JCVI rsolani.org
- *Wallemia sebi* CBS 633.66 JGI Project Link

Agaricomycetes

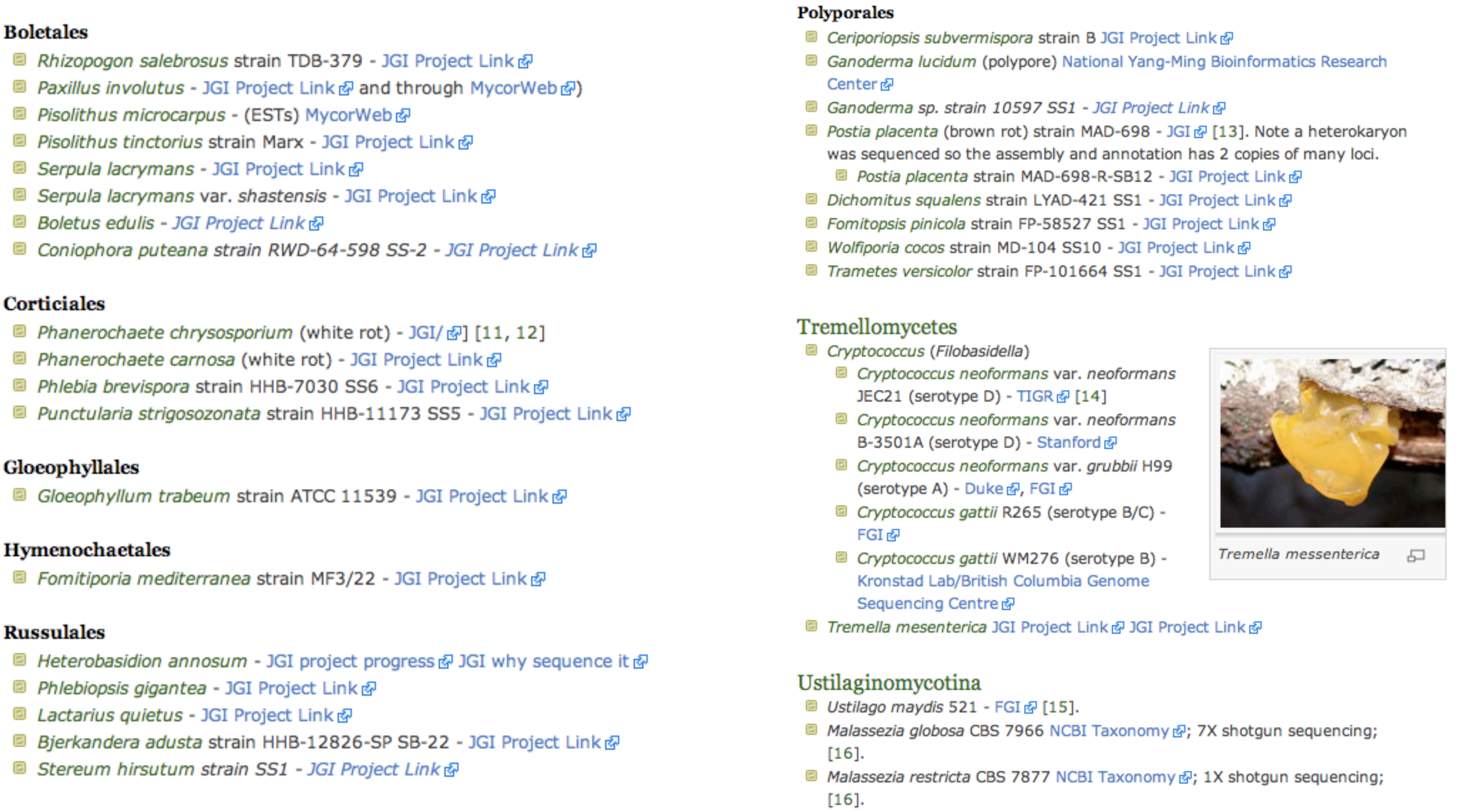
- *Auricularia delicata strain SS-5 - JGI Project Link*
- *Dacryopinax sp.* strain DJM 731 SSP-1 - JGI Project Link
- *Tomentella stuposa* - JGI Project Link

Agaricales

- *Coprinus cinereus* (*Coprinopsis cinerea*) - FGI, Duke, SEMO [7]
- *Laccaria bicolor* JGI [8]
- *Amanita bisporigera* - Michigan State/Hallen-Walton Labs 454 sequencing
- *Amanita thiersii* - JGI Project Link
- *Amanita muscaria - JGI Project Link*
- *Agaricus bisporus* var. *bisporus* strain H97 - (button mushroom) - JGI Project Link Warwick
- *Agaricus bisporus* strain var. *burnettii* JB137-S8 - JGI Project Link
- *Moniliophthora perniciosa* (causes cacao disease) (*Crinipellis perniciosa*) Unicamp Witches' Broom project page. 1.9X coverage survey sequence [9]
- *Pleurotus ostreatus* strain N001 PC15 - Announcement JGI (8X assembly)
  - strain N001 PC9 - JGI Project Link
- *Hebeloma cylindrosporum* strain H7 - JGI Project Link
  - ESTs sequenced in a separate project MycorWeb
- *Schizophyllum commune* strain H 4-8 JGI [10]
- *Piloderma croceum* strain F 1598 - JGI Project Link
- *Dacryopinax* sp. strain DJM 731 SSP-1 - JGI Project Link
- *Gloeophyllum trabeum* strain ATCC 11539 - JGI Project Link (finished)
- *Armillaria mellea - JGI Project Link*

Coprinopsis cinerea

http://fungalgenomes.org/genomes



Boletales

- *Rhizopogon salebrosus* strain TDB-379 - JGI Project Link
- *Paxillus involutus* - JGI Project Link and through MycorWeb)
- *Pisolithus microcarpus* - (ESTs) MycorWeb
- *Pisolithus tinctorius* strain Marx - JGI Project Link
- *Serpula lacrymans* - JGI Project Link
- *Serpula lacrymans* var. *shastensis* - JGI Project Link
- *Boletus edulis - JGI Project Link*
- *Coniophora puteana strain RWD-64-598 SS-2 - JGI Project Link*

Corticiales

- *Phanerochaete chrysosporium* (white rot) - JGI/] [11, 12]
- *Phanerochaete carnosa* (white rot) - JGI Project Link
- *Phlebia brevispora* strain HHB-7030 SS6 - JGI Project Link
- *Punctularia strigosozonata* strain HHB-11173 SS5 - JGI Project Link

Gloeophyllales

- *Gloeophyllum trabeum* strain ATCC 11539 - JGI Project Link

Hymenochaetales

- *Fomitiporia mediterranea* strain MF3/22 - JGI Project Link

Russulales

- *Heterobasidion annosum* - JGI project progress JGI why sequence it
- *Phlebiopsis gigantea* - JGI Project Link
- *Lactarius quietus* - JGI Project Link
- *Bjerkandera adusta* strain HHB-12826-SP SB-22 - JGI Project Link
- *Stereum hirsutum strain SS1 - JGI Project Link*

Polyporales

- *Ceriporiopsis subvermispora* strain B JGI Project Link
- *Ganoderma lucidum* (polypore) National Yang-Ming Bioinformatics Research Center
- *Ganoderma sp. strain 10597 SS1 - JGI Project Link*
- *Postia placenta* (brown rot) strain MAD-698 - JGI [13]. Note a heterokaryon was sequenced so the assembly and annotation has 2 copies of many loci.
  - *Postia placenta* strain MAD-698-R-SB12 - JGI Project Link
- *Dichomitus squalens* strain LYAD-421 SS1 - JGI Project Link
- *Fomitopsis pinicola* strain FP-58527 SS1 - JGI Project Link
- *Wolfiporia cocos* strain MD-104 SS10 - JGI Project Link
- *Trametes versicolor* strain FP-101664 SS1 - JGI Project Link

Tremellomycetes

- *Cryptococcus* (*Filobasidella*)
  - *Cryptococcus neoformans* var. *neoformans* JEC21 (serotype D) - TIGR [14]
  - *Cryptococcus neoformans* var. *neoformans* B-3501A (serotype D) - Stanford
  - *Cryptococcus neoformans* var. *grubbii* H99 (serotype A) - Duke, FGI
  - *Cryptococcus gattii* R265 (serotype B/C) - FGI
  - *Cryptococcus gattii* WM276 (serotype B) - Kronstad Lab/British Columbia Genome Sequencing Centre
- *Tremella mesenterica* JGI Project Link JGI Project Link

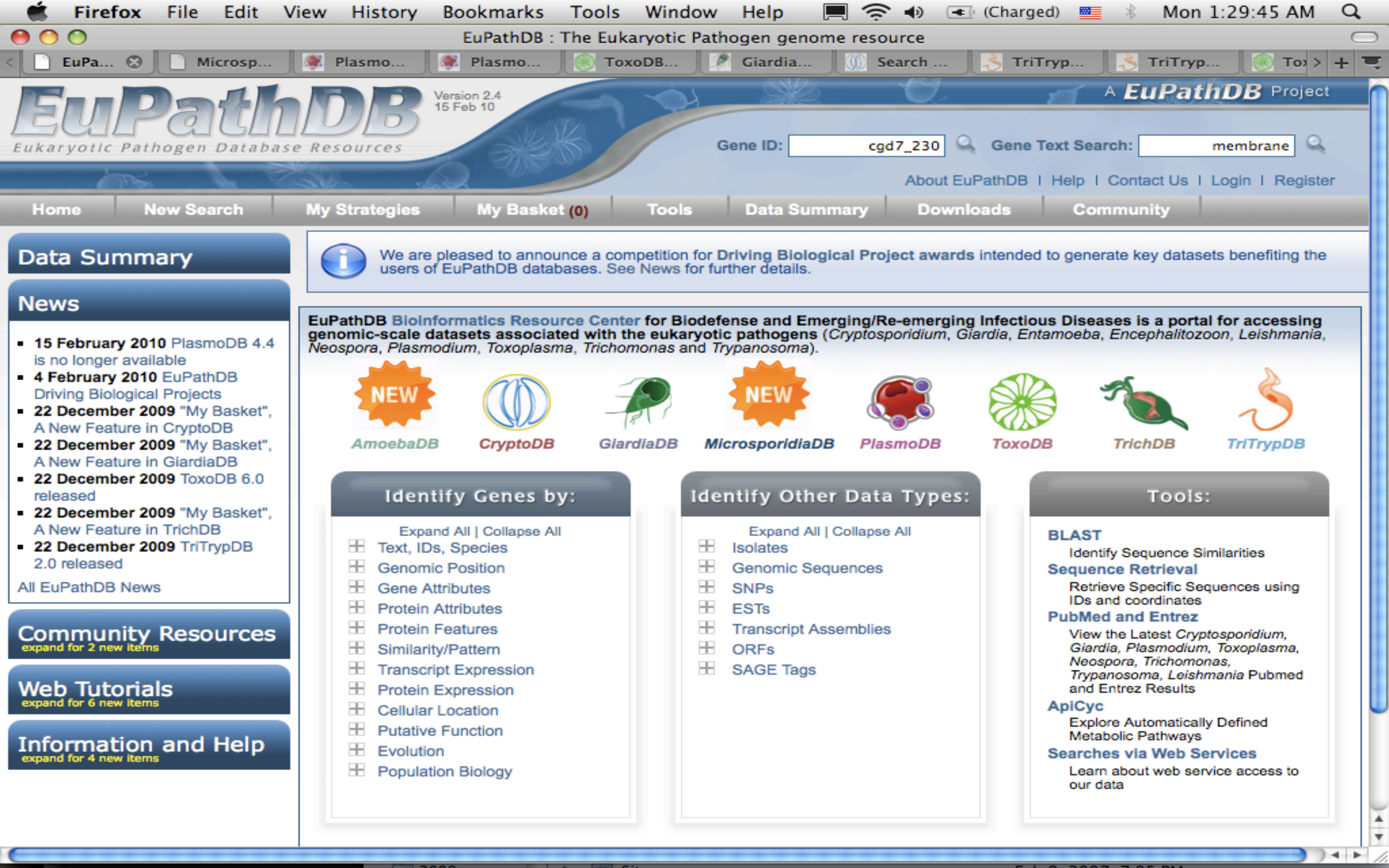
Tremella messenterica

Ustilaginomycotina

- *Ustilago maydis* 521 - FGI [15].
- *Malassezia globosa* CBS 7966 NCBI Taxonomy; 7X shotgun sequencing; [16].
- *Malassezia restricta* CBS 7877 NCBI Taxonomy; 1X shotgun sequencing; [16].

How to collect and make sense of this data?

- A few resources exist to organize what is
    - Ensembl, JGI have integrated systems for some of the data
    - Phylomedb and InParanoid dbs provides ortholog interface
- My lab is working to build pan-fungal genome database interface that is similar to resources available at http://eupathdb.org



EuPathDB : The Eukaryotic Pathogen genome resource

EuPathDB
Eukaryotic Pathogen Database Resources
Version 2.4
15 Feb 10

A **EuPathDB** Project

Gene ID: cgd7_230

Gene Text Search: membrane

Data Summary

News

- **15 February 2010** PlasmoDB 4.4 is no longer available
- **4 February 2010** EuPathDB Driving Biological Projects
- **22 December 2009** "My Basket", A New Feature in CryptoDB
- **22 December 2009** "My Basket", A New Feature in GiardiaDB
- **22 December 2009** ToxoDB 6.0 released
- **22 December 2009** "My Basket", A New Feature in TrichDB
- **22 December 2009** TriTrypDB 2.0 released

All EuPathDB News

Community Resources
expand for 2 new items

Web Tutorials
expand for 6 new items

Information and Help
expand for 4 new items

We are pleased to announce a competition for **Driving Biological Project awards** intended to generate key datasets benefiting the users of EuPathDB databases. See News for further details.

EuPathDB Bioinformatics Resource Center for Biodefense and Emerging/Re-emerging Infectious Diseases is a portal for accessing genomic-scale datasets associated with the eukaryotic pathogens (*Cryptosporidium*, *Giardia*, *Entamoeba*, *Encephalitozoon*, *Leishmania*, *Neospora*, *Plasmodium*, *Toxoplasma*, *Trichomonas* and *Trypanosoma*).

Identify Genes by:

Expand All | Collapse All

- ⊞ Text, IDs, Species
- ⊞ Genomic Position
- ⊞ Gene Attributes
- ⊞ Protein Attributes
- ⊞ Protein Features
- ⊞ Similarity/Pattern
- ⊞ Transcript Expression
- ⊞ Protein Expression
- ⊞ Cellular Location
- ⊞ Putative Function
- ⊞ Evolution
- ⊞ Population Biology

Identify Other Data Types:

Expand All | Collapse All

- ⊞ Isolates
- ⊞ Genomic Sequences
- ⊞ SNPs
- ⊞ ESTs
- ⊞ Transcript Assemblies
- ⊞ ORFs
- ⊞ SAGE Tags

Tools:

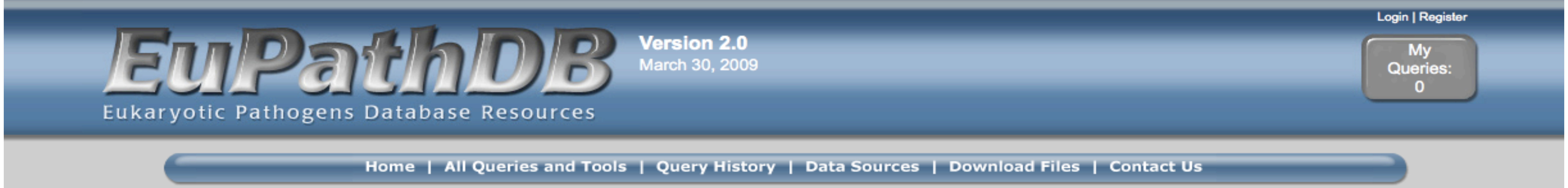
BLAST
Identify Sequence Similarities

Sequence Retrieval
Retrieve Specific Sequences using IDs and coordinates

PubMed and Entrez
View the Latest *Cryptosporidium*, *Giardia*, *Plasmodium*, *Toxoplasma*, *Neospora*, *Trichomonas*, *Trypanosoma*, *Leishmania* Pubmed and Entrez Results

ApiCyc
Explore Automatically Defined Metabolic Pathways

Searches via Web Services
Learn about web service access to our data



The philosophy behind EuPathDB databases today ...

“there’s no such thing as a stupid question”

➢ ***Curated annotation integrated with automated analyses***

predicted genes, proteins, motifs, structures, functions, pathways, *etc*

➢ ***Incorporation and integration of diverse data types***

genome & EST sequences, polymorphisms & population genetics data, chromatin modifications, transcript/protein profiling (microarrays, RNAseq, proteomics etc), interactomes, structural genomics, metabolomics, comparative genomics, reagents, field & clinical isolate phenotypes, *etc etc*

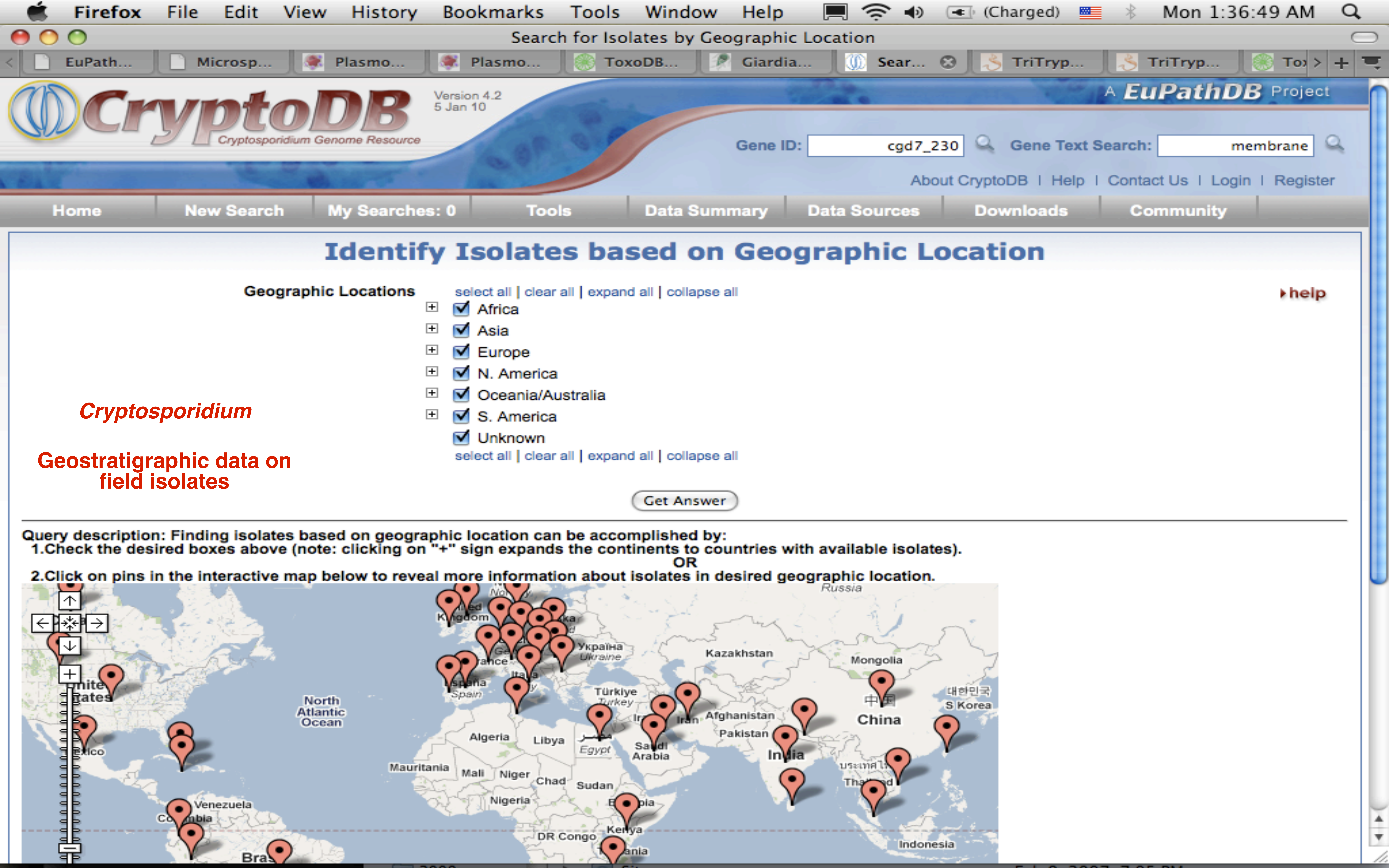
➢ ***Rapid access to both finished and unfinished (draft) data***

unfinished data is increasingly important, as genomic-scale datasets are often incomplete (or incompletely analyzed)

➢ ***Enable laboratory researchers to ask their own questions***

the goal is not an Encyclopedia of Parasitology, but hypothesis generation, providing tractable target lists amenable to validation at the bench / clinic

➢ ***Universal access.*** ~300K hits/mo, 2500+ users, 100+ countries, returning 3x/mo (ave); 3000+ citations (all EuPathDB); extensive outreach activities



Identify Isolates based on Geographic Location

Geographic Locations

select all | clear all | expand all | collapse all

- ☑ Africa
- ☑ Asia
- ☑ Europe
- ☑ N. America
- ☑ Oceania/Australia
- ☑ S. America
- ☑ Unknown

select all | clear all | expand all | collapse all

Cryptosporidium

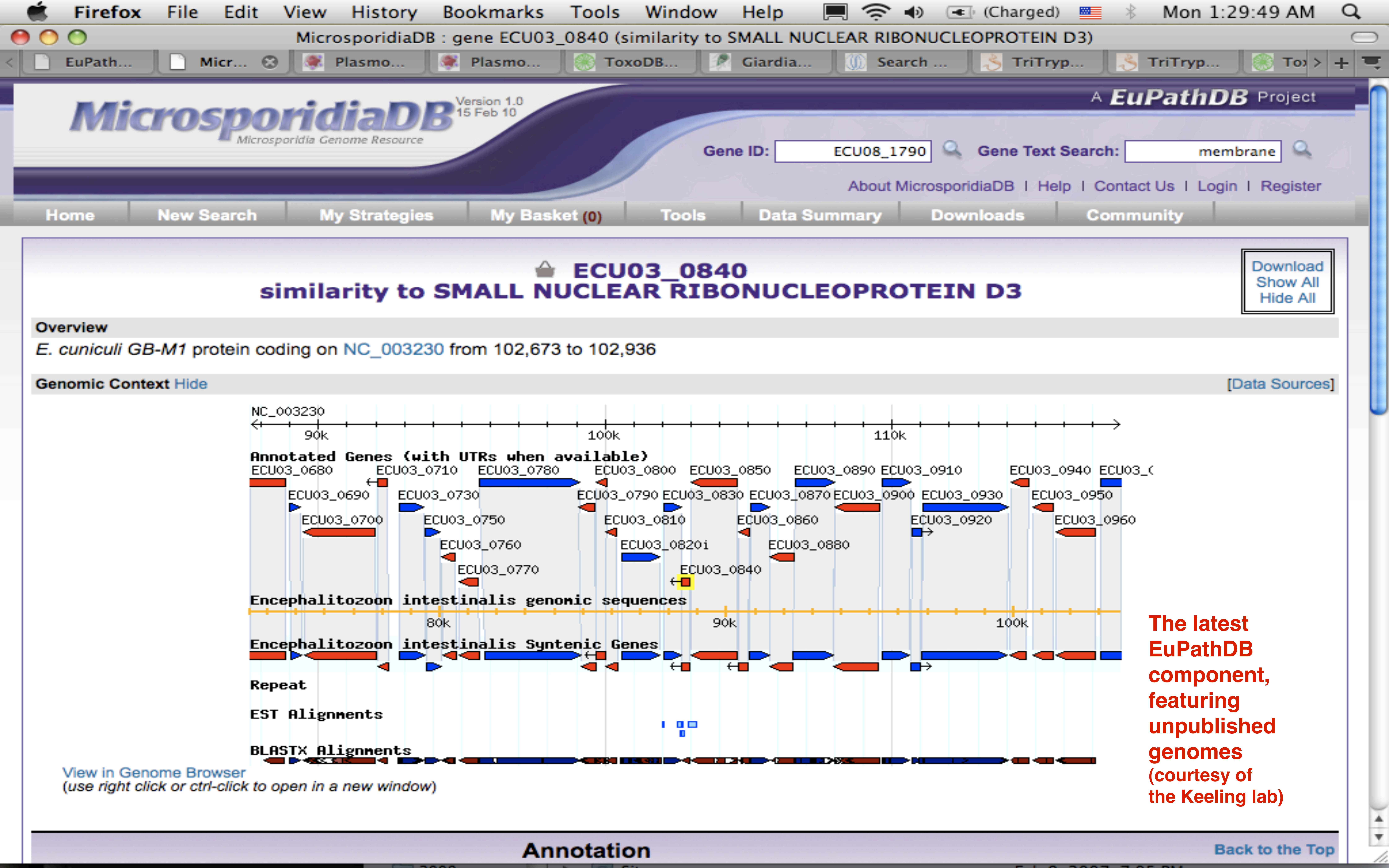
Geostratigraphic data on field isolates

Query description: Finding isolates based on geographic location can be accomplished by:

1.Check the desired boxes above (note: clicking on "+" sign expands the continents to countries with available isolates).

OR

2.Click on pins in the interactive map below to reveal more information about isolates in desired geographic location.



ECU03_0840
similarity to SMALL NUCLEAR RIBONUCLEOPROTEIN D3

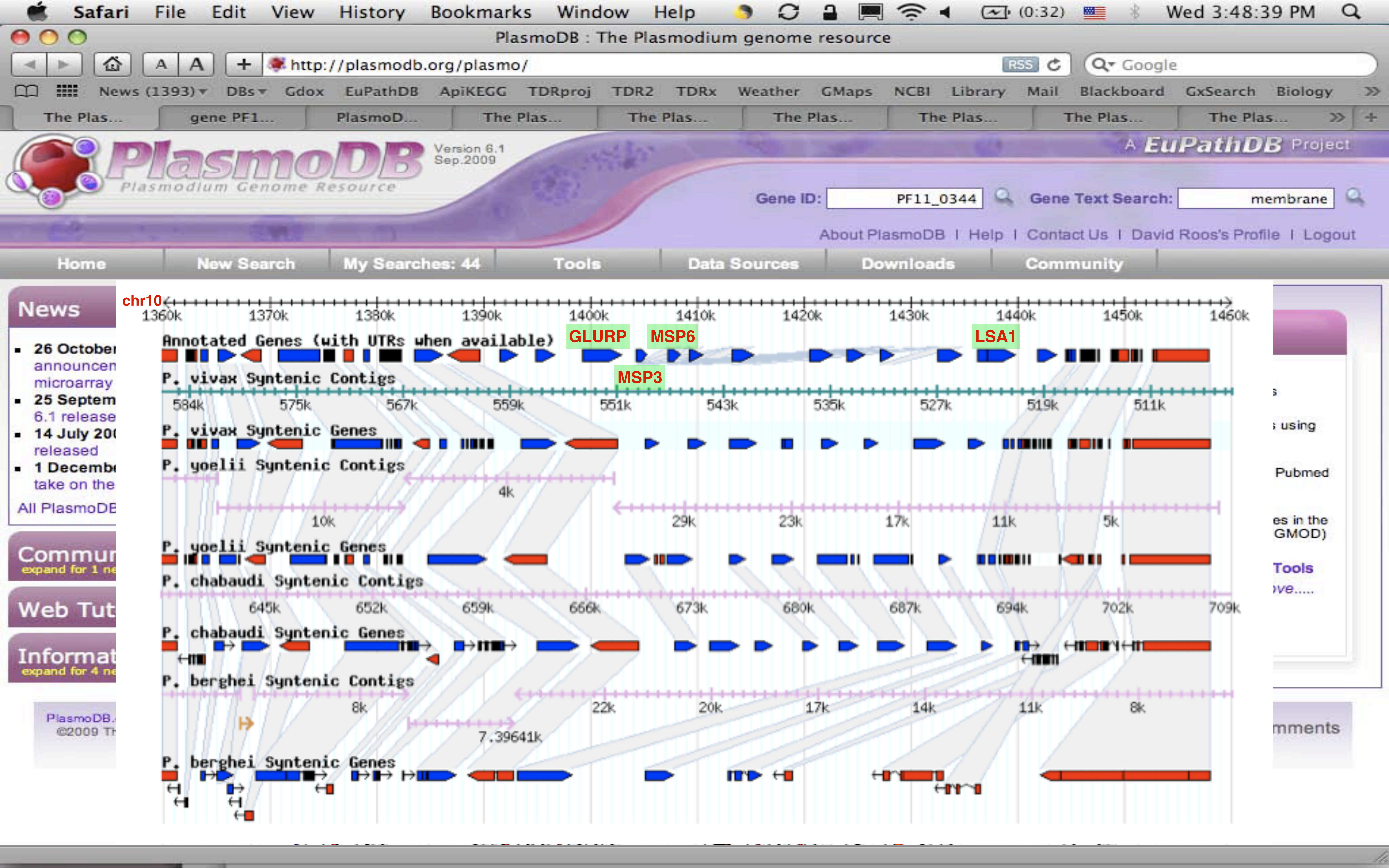
Overview

E. cuniculi GB-M1 protein coding on NC_003230 from 102,673 to 102,936

Genomic Context Hide [Data Sources]

View in Genome Browser
(use right click or ctrl-click to open in a new window)

The latest EuPathDB component, featuring unpublished genomes (courtesy of the Keeling lab)


Safari File Edit View History Bookmarks Window Help
(0:32) Wed 3:48:39 PM
PlasmoDB : The Plasmodium genome resource
http://plasmodb.org/plasmo/
RSS
Google
News (1393) DBs Gdox EuPathDB ApiKEGG TDRproj TDR2 TDRx Weather GMaps NCBI Library Mail Blackboard GxSearch Biology
The Plas... gene PF1... PlasmoD... The Plas... The Plas... The Plas... The Plas... The Plas... The Plas...
PlasmoDB
Plasmodium Genome Resource
Version 6.1
Sep.2009
A EuPathDB Project
Gene ID: PF11_0344
Gene Text Search: membrane
About PlasmoDB | Help | Contact Us | David Roos's Profile | Logout
Home
New Search
My Searches: 44
Tools
Data Sources
Downloads
Community
News
26 Octobe
announcen
microarray
25 Septem
6.1 release
14 July 20
released
1 Decemb
take on the
All PlasmoDE
Commun
expand for 1 ne
Web Tut
Informat
expand for 4 ne
PlasmoDB.
©2009 Th
s
using
Pubmed
es in the
GMOD)
Tools
ive.....
mments
chr10
1360k 1370k 1380k 1390k 1400k 1410k 1420k 1430k 1440k 1450k 1460k
Annotated Genes (with UTRs when available)
GLURP
MSP6
LSA1
P. vivax Syntenic Contigs
MSP3
584k 575k 567k 559k 551k 543k 535k 527k 519k 511k
P. vivax Syntenic Genes
P. yoelii Syntenic Contigs
4k
10k 29k 23k 17k 11k 5k
P. yoelii Syntenic Genes
P. chabaudi Syntenic Contigs
645k 652k 659k 666k 673k 680k 687k 694k 702k 709k
P. chabaudi Syntenic Genes
P. berghei Syntenic Contigs
8k 22k 20k 17k 14k 11k 8k
7.39641k
P. berghei Syntenic Genes

What genomes are being generated and where?

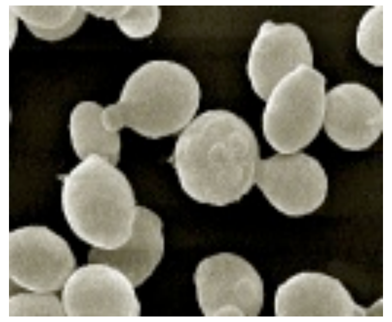
Bioenergy Prospective on Fungi

Lignocellulose degraders

Grow → Degrade → Ferment

Plant symbionts and pathogens

Yeasts



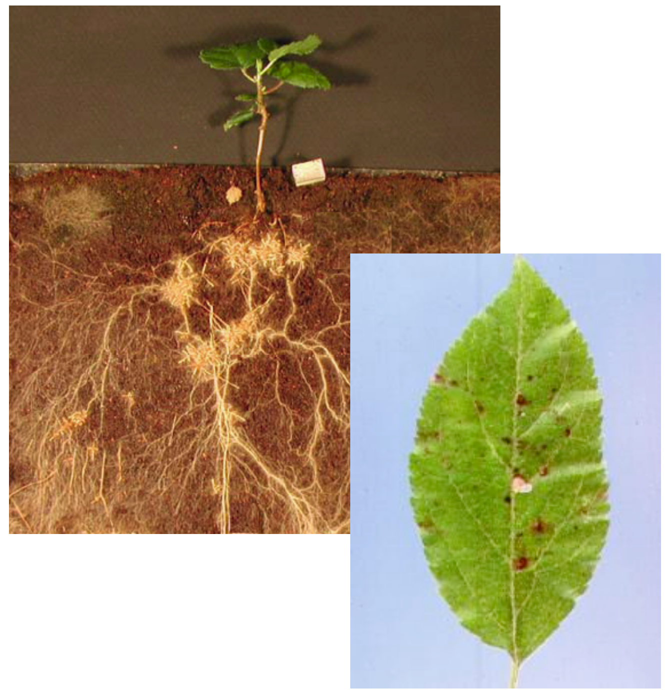
Fungal Genomics Program Goals

- Develop systematic approaches and scale up sequencing and analysis of fungal genomes.
- Explore phylogenetic and ecological diversity of fungi for DOE relevant science and applications

Genomic Encyclopedia of Fungi

Vol.1. **Plant feedstock health**

Part 1. Mycorrhizal Symbiosis
Part 2. Plant Pathogenicity
Part 3. Biocontrol

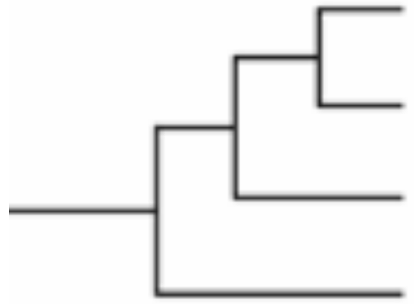


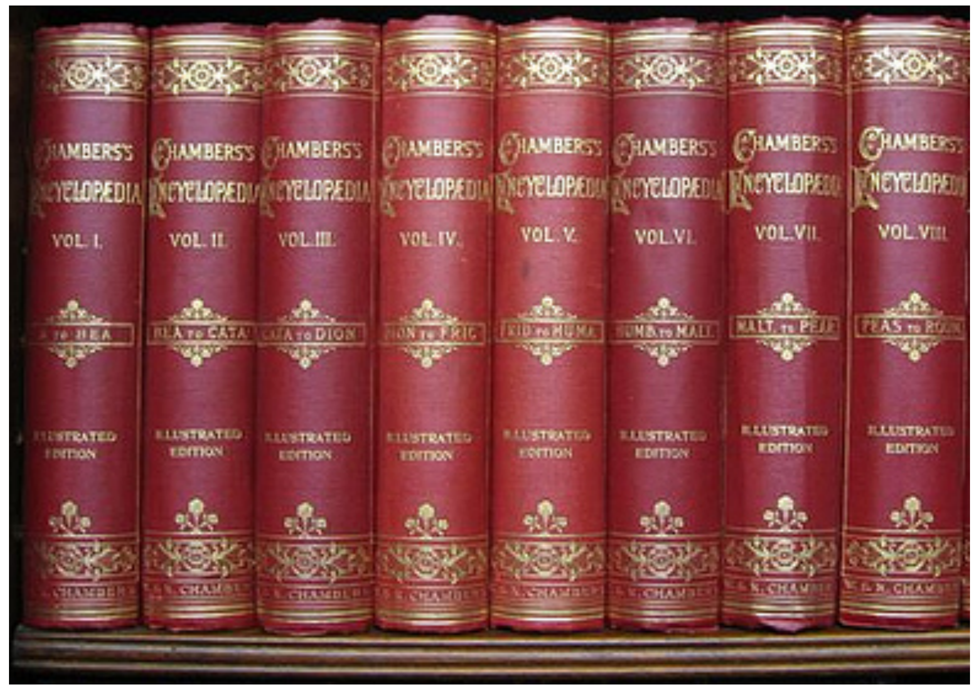
Vol.2. **Biorefinery fungi**

Part 1. Lignocellulose degradation
Part 2. Sugar fermentation
Part 3. Industrial organisms

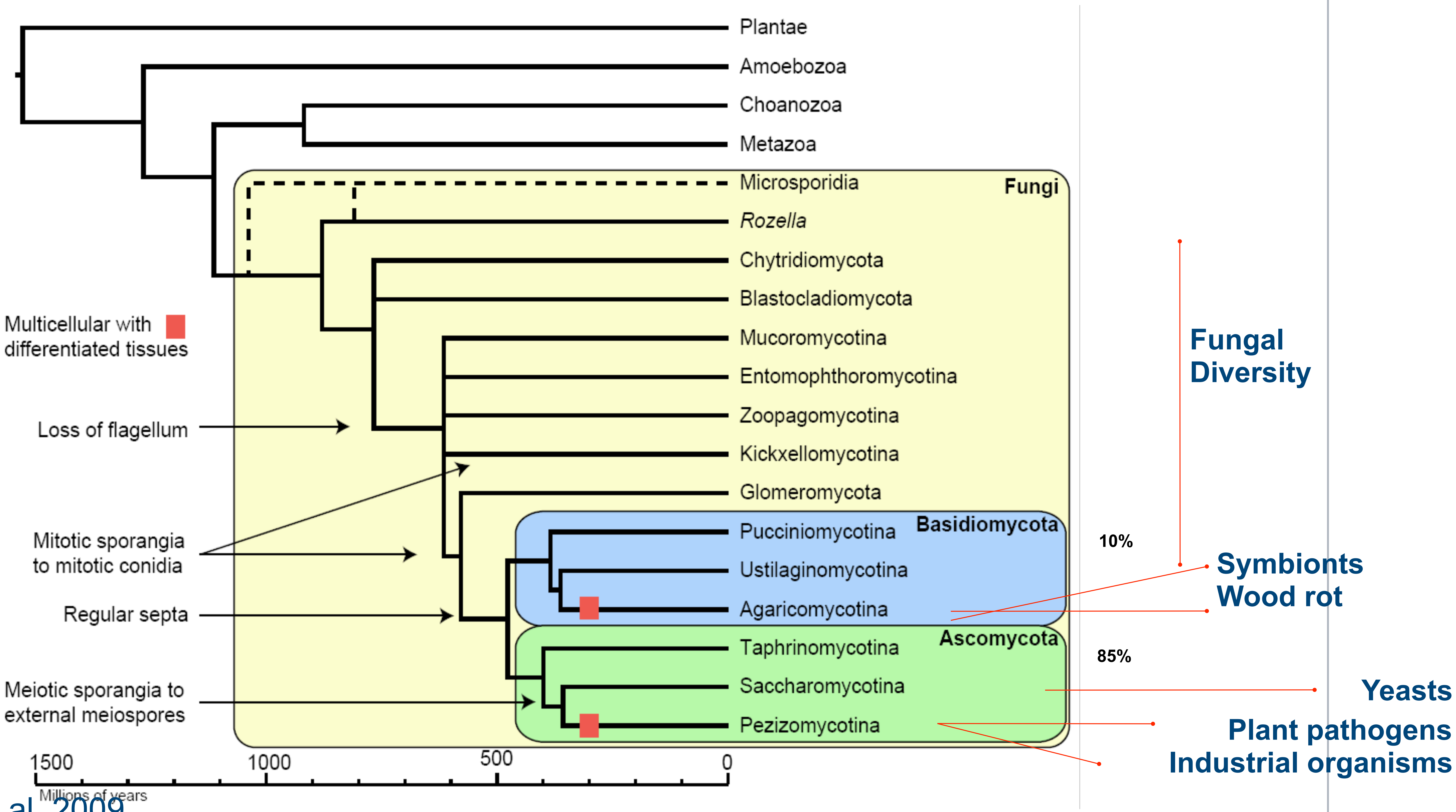
Vol.3. **Fungal diversity**

Part 1. Phylogenetic diversity
Part 2. Ecological diversity

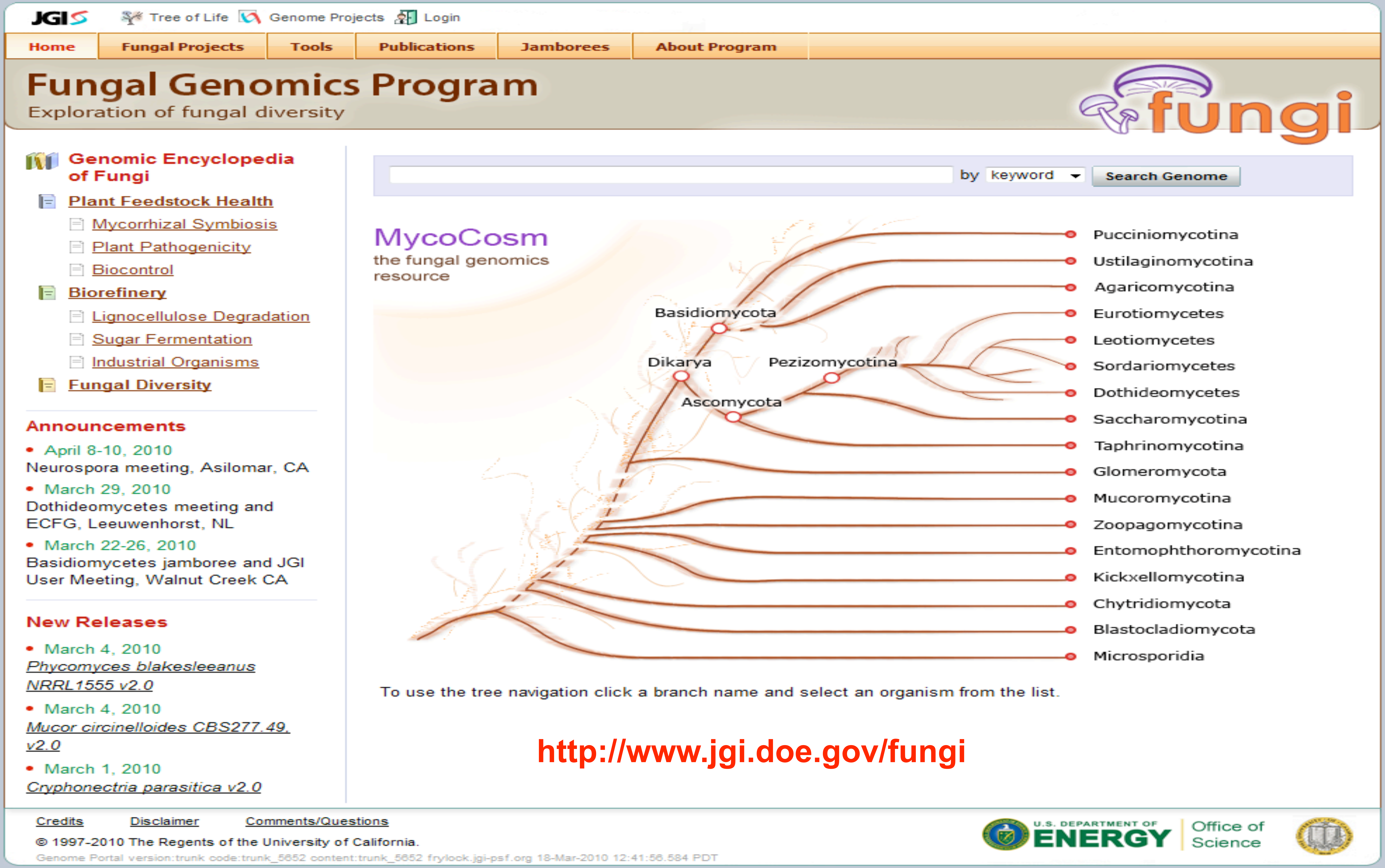




Reading Fungi at Different Resolution



Stajich et al, 2009



Fungal Genomics Program

Exploration of fungal diversity

Genomic Encyclopedia of Fungi

- Plant Feedstock Health
  - Mycorrhizal Symbiosis
  - Plant Pathogenicity
  - Biocontrol
- Biorefinery
  - Lignocellulose Degradation
  - Sugar Fermentation
  - Industrial Organisms
- Fungal Diversity

Announcements

- April 8-10, 2010
  Neurospora meeting, Asilomar, CA
- March 29, 2010
  Dothideomycetes meeting and ECFG, Leeuwenhorst, NL
- March 22-26, 2010
  Basidiomycetes jamboree and JGI User Meeting, Walnut Creek CA

New Releases

- March 4, 2010
  Phycomyces blakesleeanus NRRL1555 v2.0
- March 4, 2010
  Mucor circinelloides CBS277.49, v2.0
- March 1, 2010
  Cryphonectria parasitica v2.0

MycoCosm

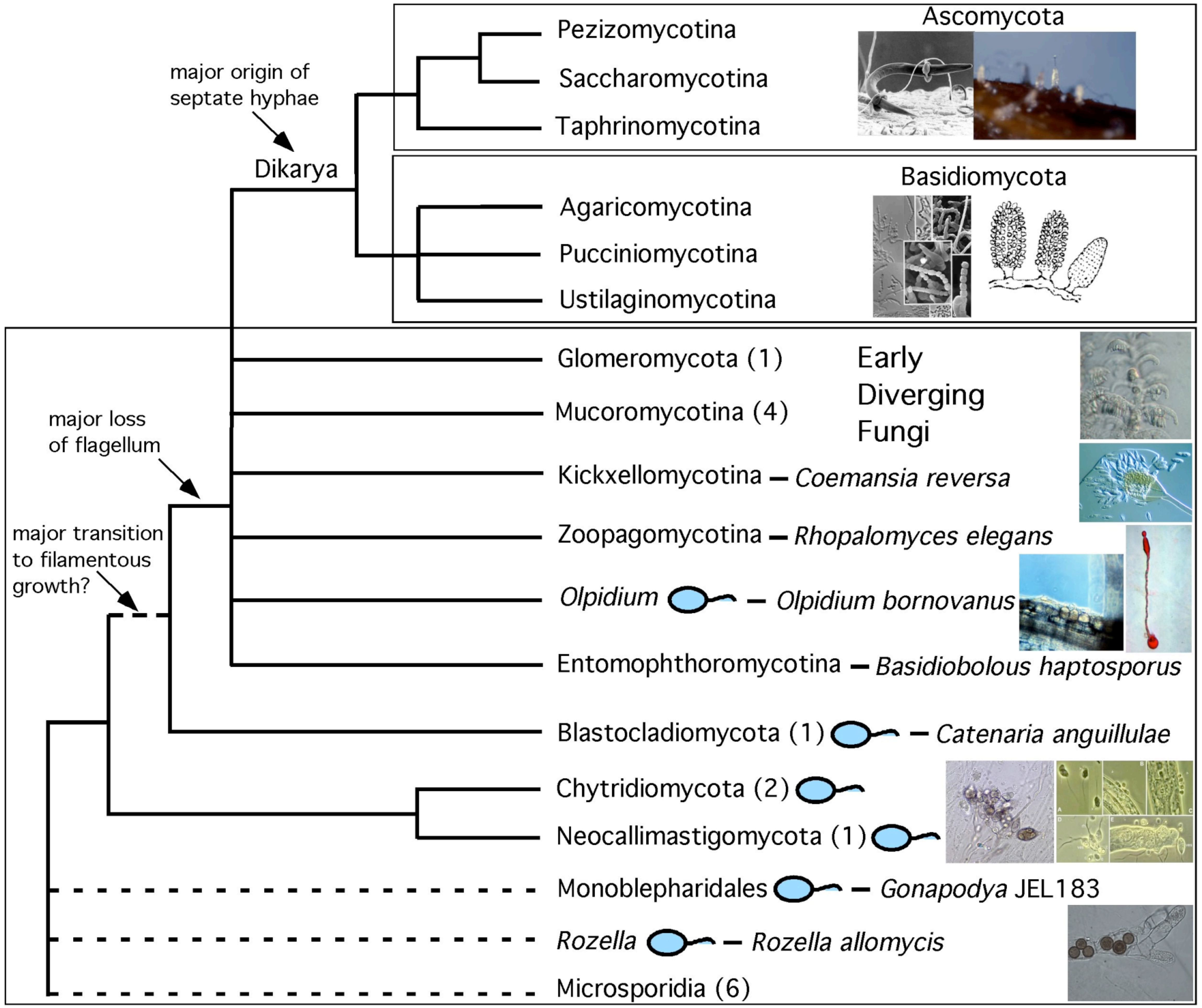
the fungal genomics resource

To use the tree navigation click a branch name and select an organism from the list.

http://www.jgi.doe.gov/fungi

Genome Portal version:trunk code:trunk_5652 content:trunk_5652 frylock.jgi-psf.org 18-Mar-2010 12:41:56.584 PDT

Proposed Targets for JGI Fungal Genomics Project - Diversity

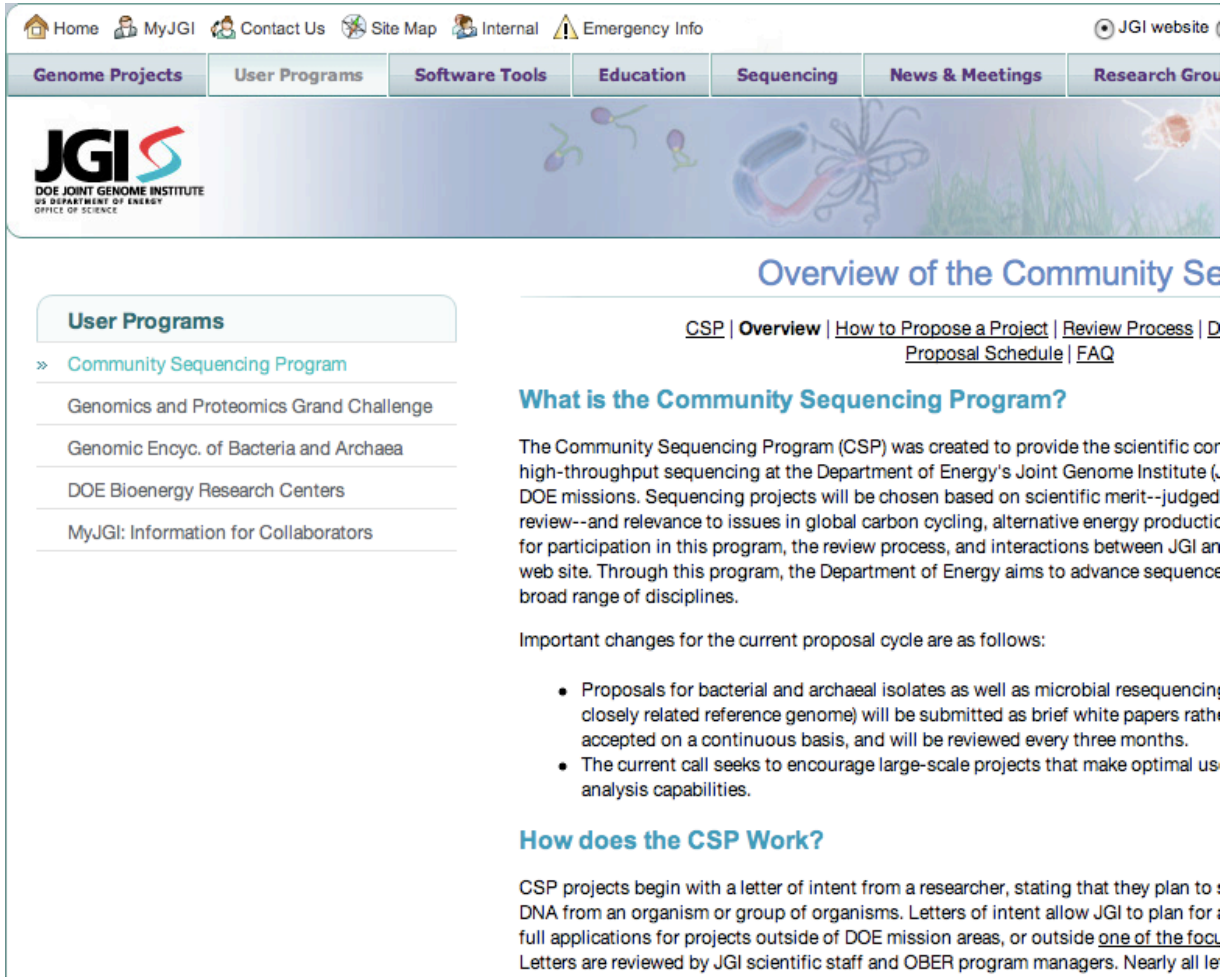


JGI FGP Project foci - proposals

- Diversity
- Fermentation - Yeasts
- Growth - symbiotic genomics of mainly basidiomycete
- Feed stock - Dothideomycetes
- Cellulosic Degradation - Mainly basidiomycetes. Enzyme discovery

You can submit your own proposals

- Community Sequencing Program for DOE relevant proposals and how they address “global carbon cycling, alternative energy production, and biogeochemistry.”
    - Previously bioremediation, radiation resistance, etc.
    - Typically May deadline each year with a preproposal deadline in March.
    - No funds for you but will sequence and annotate the genome and provides tools for community annotation and participation



Overview of the Community Se

What is the Community Sequencing Program?

The Community Sequencing Program (CSP) was created to provide the scientific cor
high-throughput sequencing at the Department of Energy's Joint Genome Institute (
DOE missions. Sequencing projects will be chosen based on scientific merit--judged
review--and relevance to issues in global carbon cycling, alternative energy productic
for participation in this program, the review process, and interactions between JGI an
web site. Through this program, the Department of Energy aims to advance sequence
broad range of disciplines.

Important changes for the current proposal cycle are as follows:

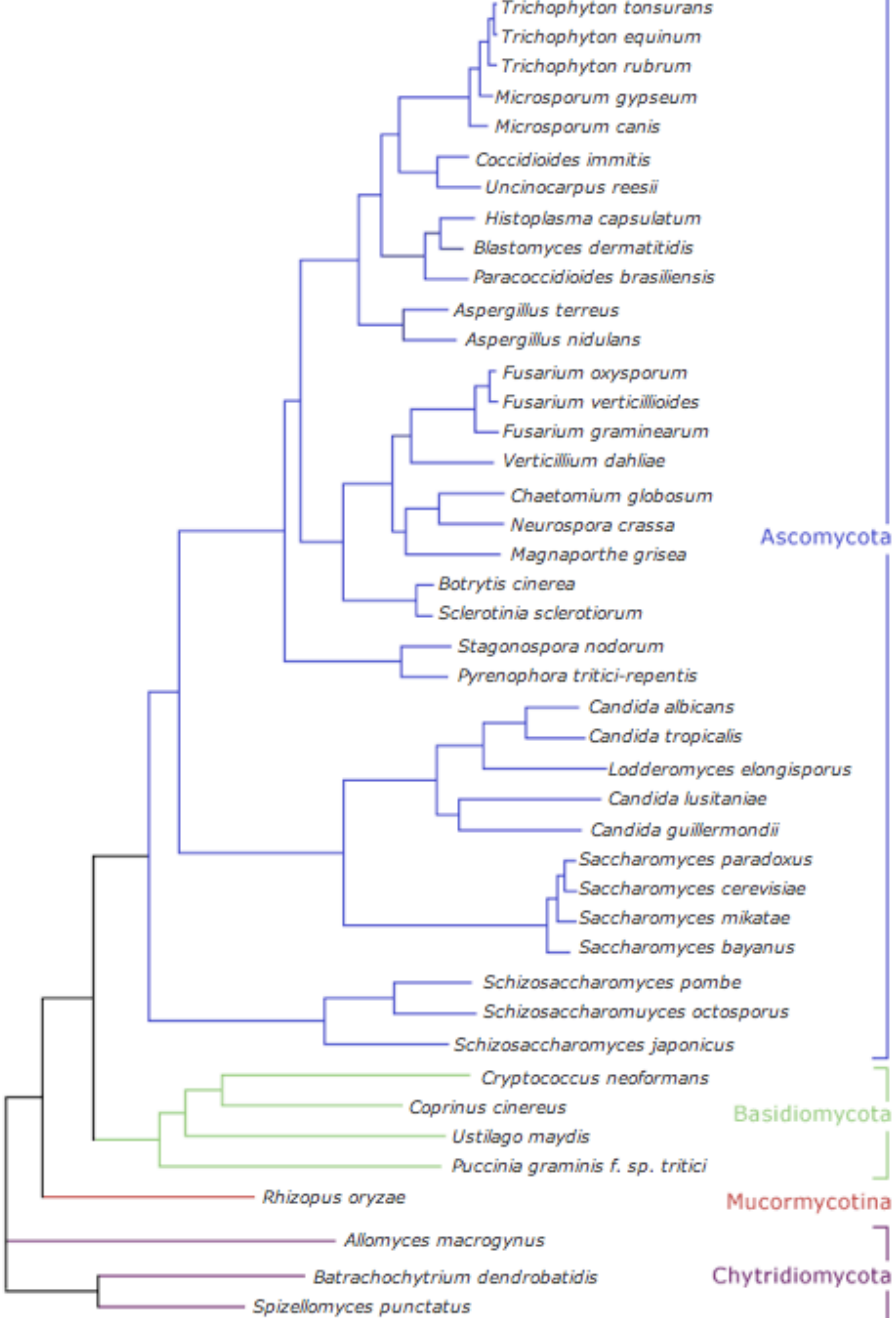
- Proposals for bacterial and archaeal isolates as well as microbial resequencin
  closely related reference genome) will be submitted as brief white papers rath
  accepted on a continuous basis, and will be reviewed every three months.
- The current call seeks to encourage large-scale projects that make optimal us
  analysis capabilities.

How does the CSP Work?

CSP projects begin with a letter of intent from a researcher, stating that they plan to
DNA from an organism or group of organisms. Letters of intent allow JGI to plan for
full applications for projects outside of DOE mission areas, or outside one of the focu
Letters are reviewed by JGI scientific staff and OBER program managers. Nearly all le

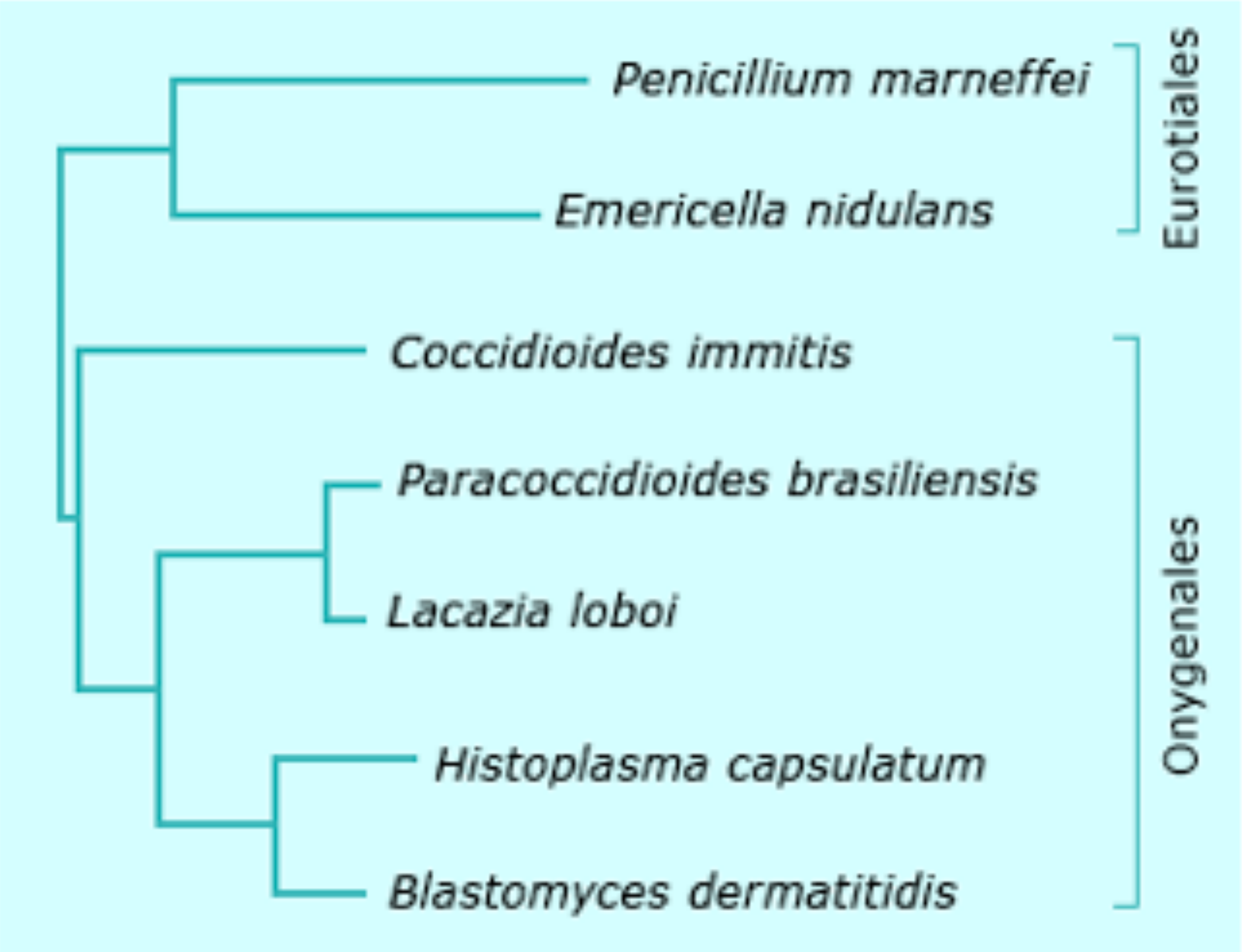
Broad Institute

- Fungal Genome Initiative
- Model systems and then Human health focused
- Human microbiome project
  - Dermatophyte fungi
- Also other foci that relate to human health, e.g. vector biology (A. or some crop diseases *P. infestans*

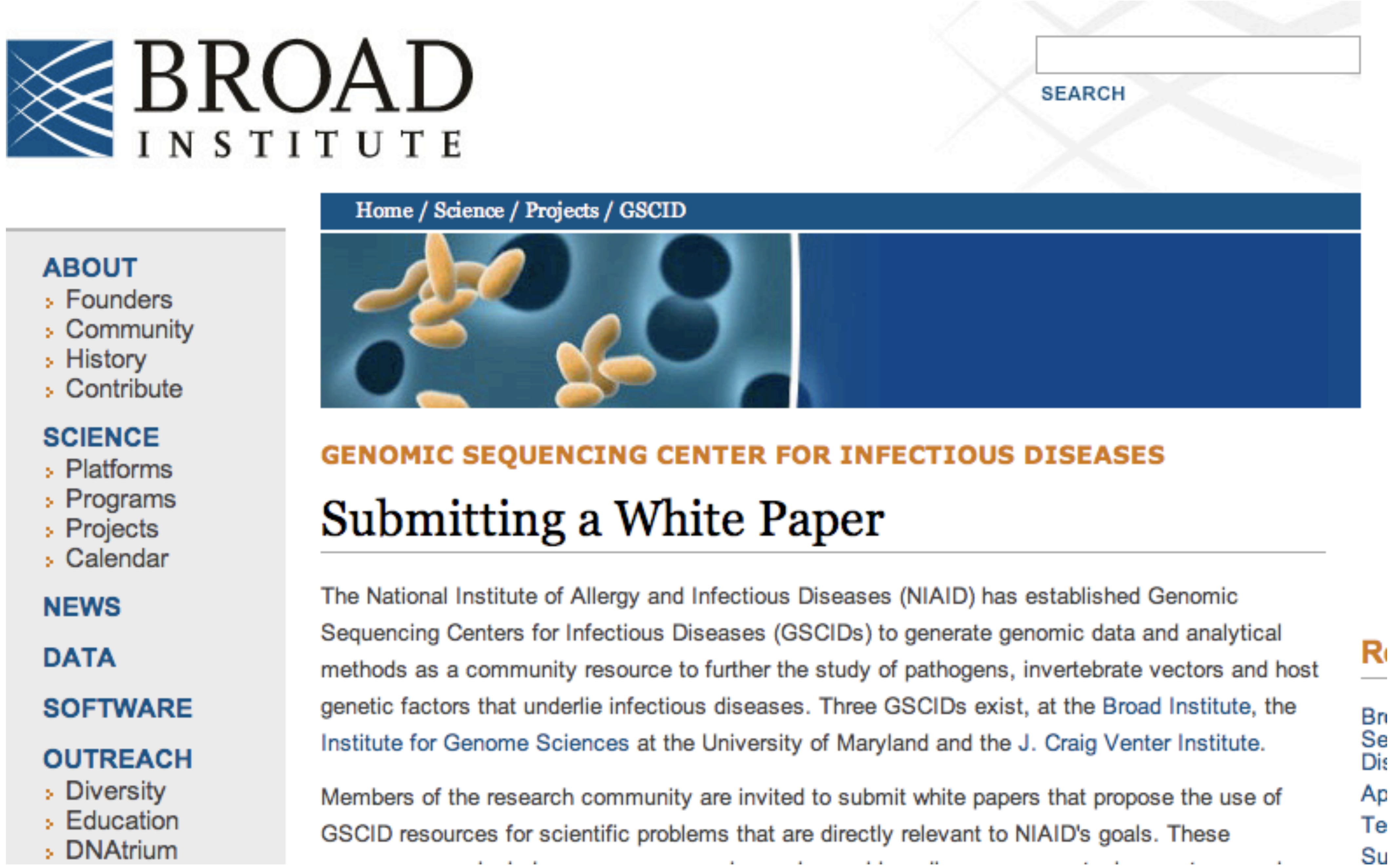


Other foci

- Dimorphic fungi - Onygenales fungi
  - 14 genomes of *Coccidioides* strains (*C. immitis* and *C. posadasii*)
- Can submit your own proposals to MSC for infectious diseases



Neighbor-joining tree of dimorphic fungi. Two Chaetotyriales species were used as outgroup taxa (modified from Herr et al. 2001).



GENOMIC SEQUENCING CENTER FOR INFECTIOUS DISEASES

Submitting a White Paper

The National Institute of Allergy and Infectious Diseases (NIAID) has established Genomic Sequencing Centers for Infectious Diseases (GSCIDs) to generate genomic data and analytical methods as a community resource to further the study of pathogens, invertebrate vectors and host genetic factors that underlie infectious diseases. Three GSCIDs exist, at the Broad Institute, the Institute for Genome Sciences at the University of Maryland and the J. Craig Venter Institute.

Members of the research community are invited to submit white papers that propose the use of GSCID resources for scientific problems that are directly relevant to NIAID's goals. These

Other major centers and projects

- Genolevures project generated 6 high quality assemblies and annotations among Saccharomycotina.
- Genoscope DIKARYOME project will generate yeast genome project sequencing dozens of species with yeast forms from Ascomycota and Basidiomycota.
- Washington University, St Louis - *Histoplasma*, *Alternaria*, 25 strains of *S. cerevisiae*
- Sanger Institute - had completed several fungi (*S. pombe*, *A. fumigatus*)
- BGI in China - dozens of fungi

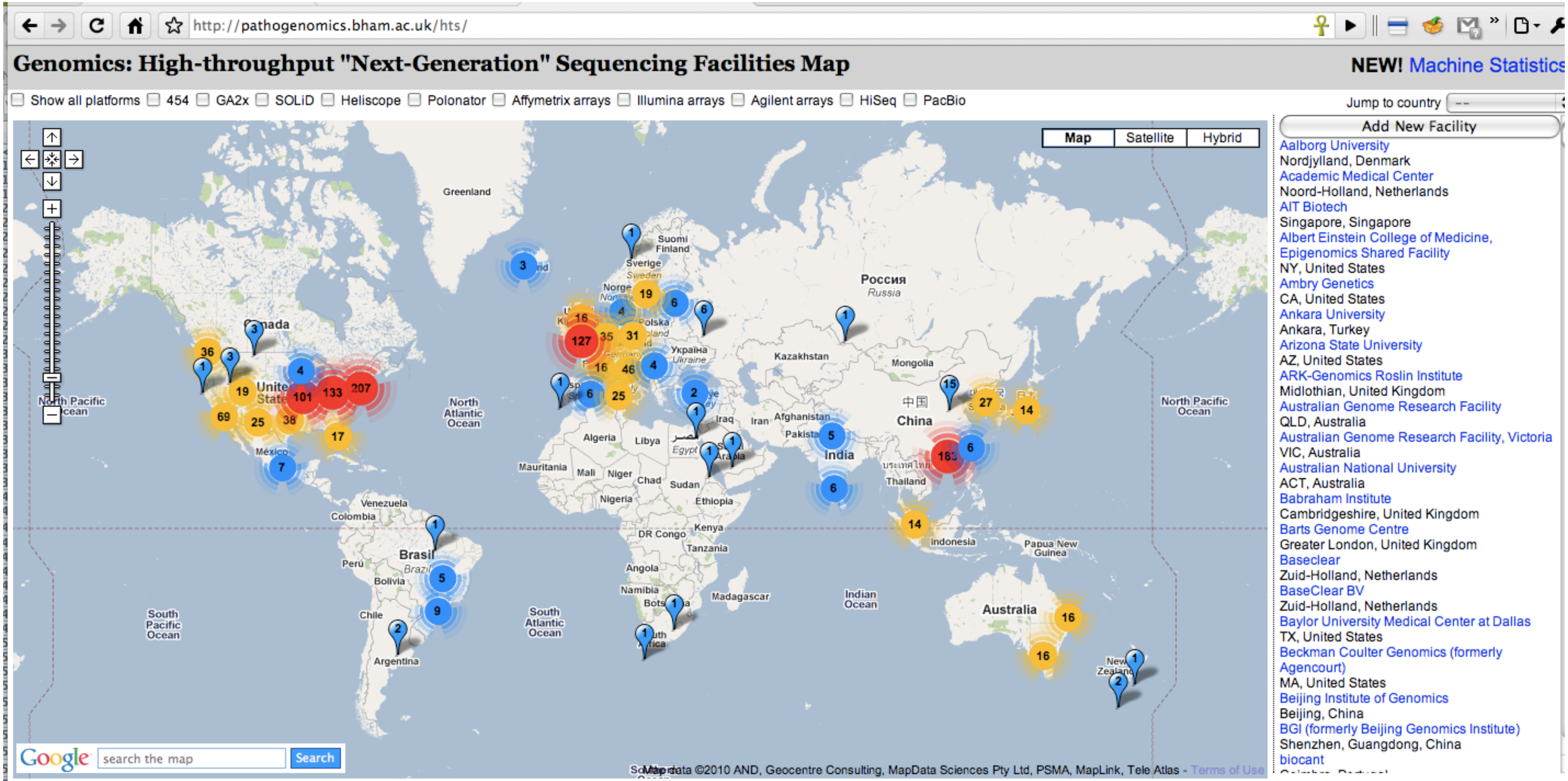
DIY?

Should you be sequencing those genomes yourself instead?

Strategies for sequencing

- Academic sequencing centers
  - Rates are better - queue may be longer.
  - Illumina GAII - ~USD$1000-1800 for single-ended (36-72 bp);
    $3000 for 76bp x Paired-End
    20-25M Tags = for 76bp PE = 3.8 Gb for $3k = ~100X coverage of 40Mb genome!
- Commercial options abound now as well.
- Reality is you are going to need Bioinformatics support for anything.
- Mixed size insert libraries and read lengths are probably needed to cover repeats

Many places to choose from (1232 machines in 397 centers as of 1-August-2010)



Genomics: High-throughput "Next-Generation" Sequencing Facilities Map

NEW! Machine Statistics

☐ Show all platforms ☐ 454 ☐ GA2x ☐ SOLiD ☐ Heliscope ☐ Polonator ☐ Affymetrix arrays ☐ Illumina arrays ☐ Agilent arrays ☐ HiSeq ☐ PacBio

Jump to country --

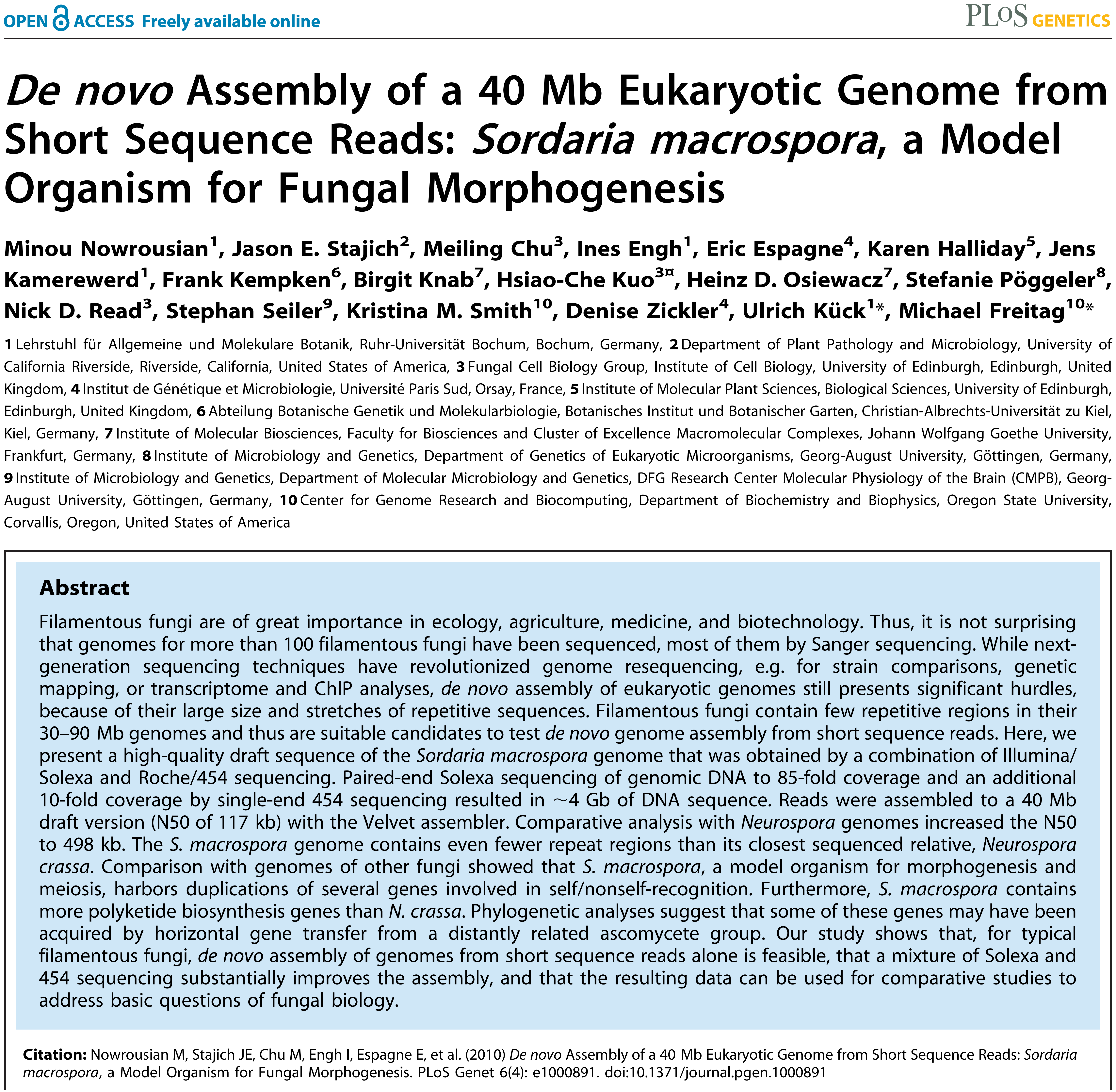
Aalborg University
Nordjylland, Denmark
Academic Medical Center
Noord-Holland, Netherlands
AIT Biotech
Singapore, Singapore
Albert Einstein College of Medicine, Epigenomics Shared Facility
NY, United States
Ambry Genetics
CA, United States
Ankara University
Ankara, Turkey
Arizona State University
AZ, United States
ARK-Genomics Roslin Institute
Midlothian, United Kingdom
Australian Genome Research Facility
QLD, Australia
Australian Genome Research Facility, Victoria
VIC, Australia
Australian National University
ACT, Australia
Babraham Institute
Cambridgeshire, United Kingdom
Barts Genome Centre
Greater London, United Kingdom
Baseclear
Zuid-Holland, Netherlands
BaseClear BV
Zuid-Holland, Netherlands
Baylor University Medical Center at Dallas
TX, United States
Beckman Coulter Genomics (formerly Agencourt)
MA, United States
Beijing Institute of Genomics
Beijing, China
BGI (formerly Beijing Genomics Institute)
Shenzhen, Guangdong, China
biocant

Map data ©2010 AND, Geocentre Consulting, MapData Sciences Pty Ltd, PSMA, MapLink, Tele Atlas

The hybrid model

- Mix reads from different technologies - 454 & Illumina
- Probably limits the biases from different technologies but also requires some additional work for assembly
- Mixed library insert size and paired-end data can make all the difference in success
- Close relative makes scaffolding possible.

1 year genome project!



OPEN ACCESS Freely available online

PLoS GENETICS

De novo Assembly of a 40 Mb Eukaryotic Genome from Short Sequence Reads: *Sordaria macrospora*, a Model Organism for Fungal Morphogenesis

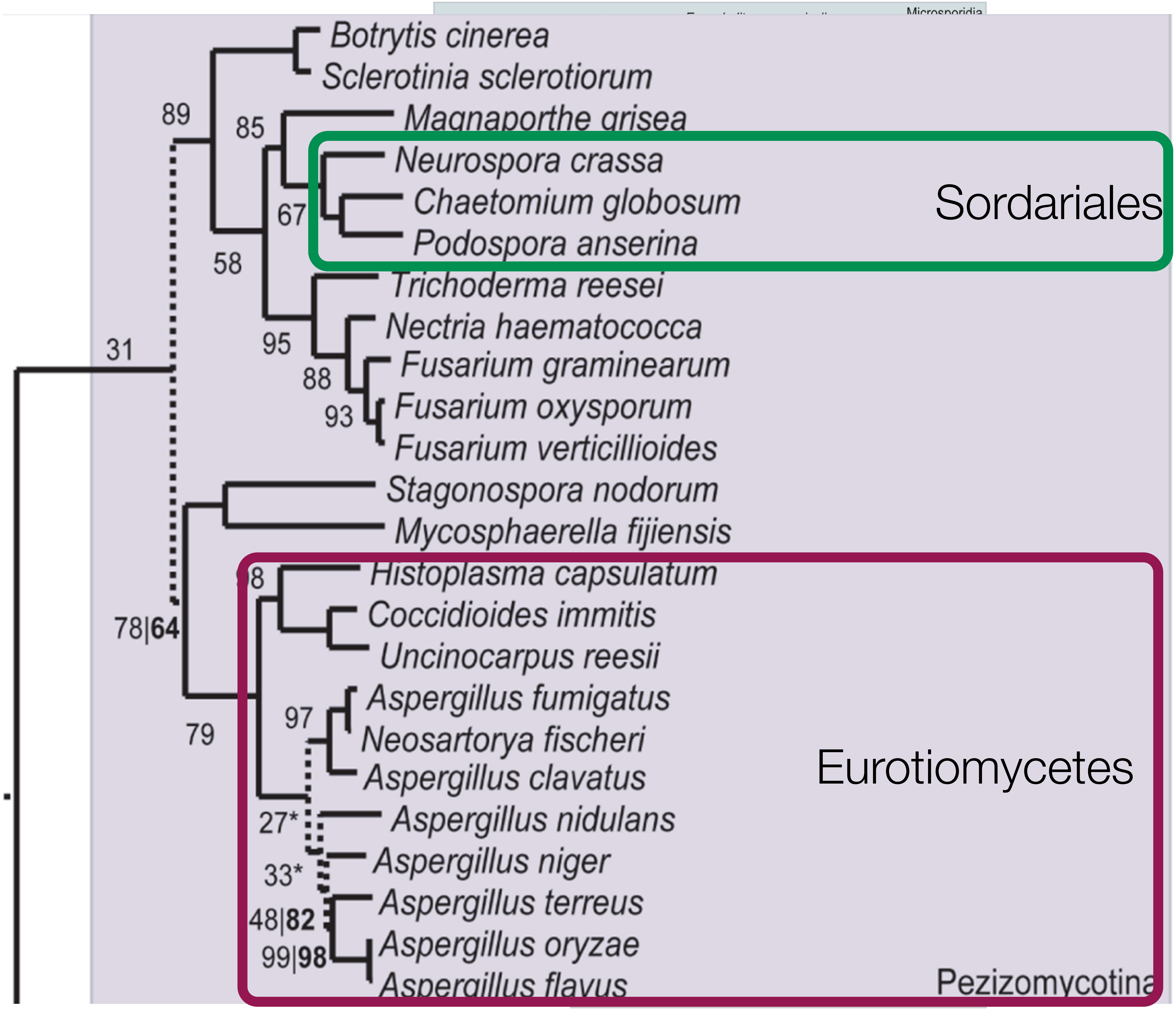
Minou Nowrousian[1], Jason E. Stajich[2], Meiling Chu[3], Ines Engh[1], Eric Espagne[4], Karen Halliday[5], Jens Kamerewerd[1], Frank Kempken[6], Birgit Knab[7], Hsiao-Che Kuo[3¤], Heinz D. Osiewacz[7], Stefanie Pöggeler[8], Nick D. Read[3], Stephan Seiler[9], Kristina M. Smith[10], Denise Zickler[4], Ulrich Kück[1*], Michael Freitag[10*]

1 Lehrstuhl für Allgemeine und Molekulare Botanik, Ruhr-Universität Bochum, Bochum, Germany, **2** Department of Plant Pathology and Microbiology, University of California Riverside, Riverside, California, United States of America, **3** Fungal Cell Biology Group, Institute of Cell Biology, University of Edinburgh, Edinburgh, United Kingdom, **4** Institut de Génétique et Microbiologie, Université Paris Sud, Orsay, France, **5** Institute of Molecular Plant Sciences, Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom, **6** Abteilung Botanische Genetik und Molekularbiologie, Botanisches Institut und Botanischer Garten, Christian-Albrechts-Universität zu Kiel, Kiel, Germany, **7** Institute of Molecular Biosciences, Faculty for Biosciences and Cluster of Excellence Macromolecular Complexes, Johann Wolfgang Goethe University, Frankfurt, Germany, **8** Institute of Microbiology and Genetics, Department of Genetics of Eukaryotic Microorganisms, Georg-August University, Göttingen, Germany, **9** Institute of Microbiology and Genetics, Department of Molecular Microbiology and Genetics, DFG Research Center Molecular Physiology of the Brain (CMPB), Georg-August University, Göttingen, Germany, **10** Center for Genome Research and Biocomputing, Department of Biochemistry and Biophysics, Oregon State University, Corvallis, Oregon, United States of America

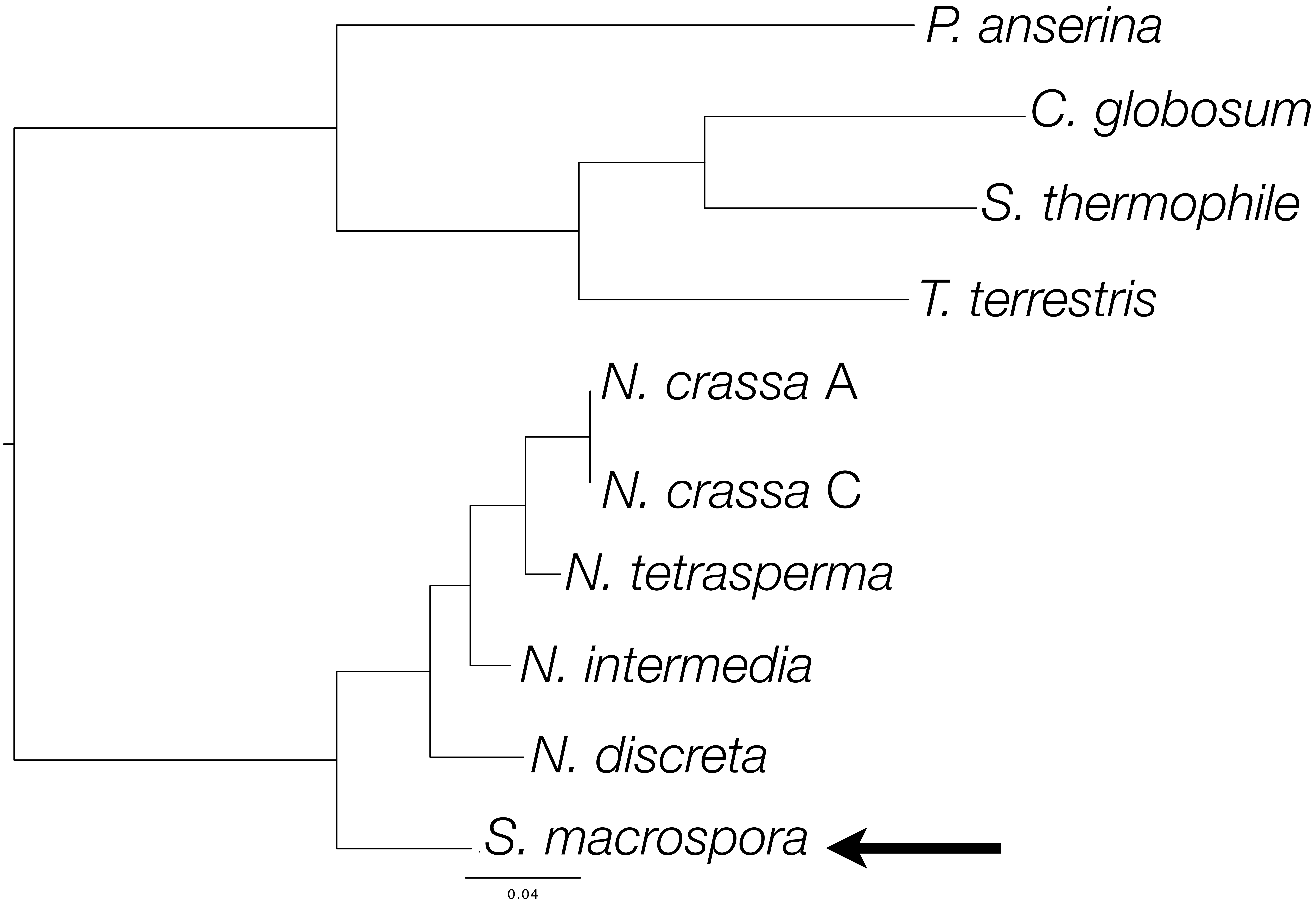
Abstract

Filamentous fungi are of great importance in ecology, agriculture, medicine, and biotechnology. Thus, it is not surprising that genomes for more than 100 filamentous fungi have been sequenced, most of them by Sanger sequencing. While next-generation sequencing techniques have revolutionized genome resequencing, e.g. for strain comparisons, genetic mapping, or transcriptome and ChIP analyses, *de novo* assembly of eukaryotic genomes still presents significant hurdles, because of their large size and stretches of repetitive sequences. Filamentous fungi contain few repetitive regions in their 30–90 Mb genomes and thus are suitable candidates to test *de novo* genome assembly from short sequence reads. Here, we present a high-quality draft sequence of the *Sordaria macrospora* genome that was obtained by a combination of Illumina/Solexa and Roche/454 sequencing. Paired-end Solexa sequencing of genomic DNA to 85-fold coverage and an additional 10-fold coverage by single-end 454 sequencing resulted in ~4 Gb of DNA sequence. Reads were assembled to a 40 Mb draft version (N50 of 117 kb) with the Velvet assembler. Comparative analysis with *Neurospora* genomes increased the N50 to 498 kb. The *S. macrospora* genome contains even fewer repeat regions than its closest sequenced relative, *Neurospora crassa*. Comparison with genomes of other fungi showed that *S. macrospora*, a model organism for morphogenesis and meiosis, harbors duplications of several genes involved in self/nonself-recognition. Furthermore, *S. macrospora* contains more polyketide biosynthesis genes than *N. crassa*. Phylogenetic analyses suggest that some of these genes may have been acquired by horizontal gene transfer from a distantly related ascomycete group. Our study shows that, for typical filamentous fungi, *de novo* assembly of genomes from short sequence reads alone is feasible, that a mixture of Solexa and 454 sequencing substantially improves the assembly, and that the resulting data can be used for comparative studies to address basic questions of fungal biology.

Citation: Nowrousian M, Stajich JE, Chu M, Engh I, Espagne E, et al. (2010) *De novo* Assembly of a 40 Mb Eukaryotic Genome from Short Sequence Reads: *Sordaria macrospora*, a Model Organism for Fungal Morphogenesis. PLoS Genet 6(4): e1000891. doi:10.1371/journal.pgen.1000891

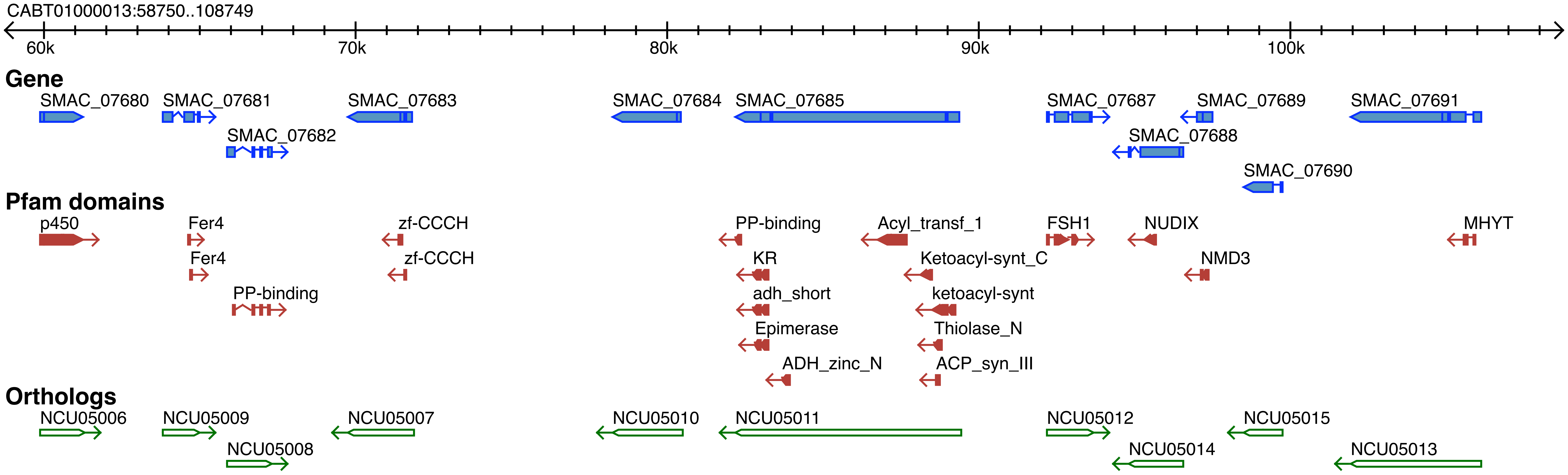


ıbe and Gabaldón
One 2009.


P. anserina
C. globosum
S. thermophile
T. terrestris
N. crassa A
N. crassa C
N. tetrasperma
N. intermedia
N. discreta
S. macrospora
0.04

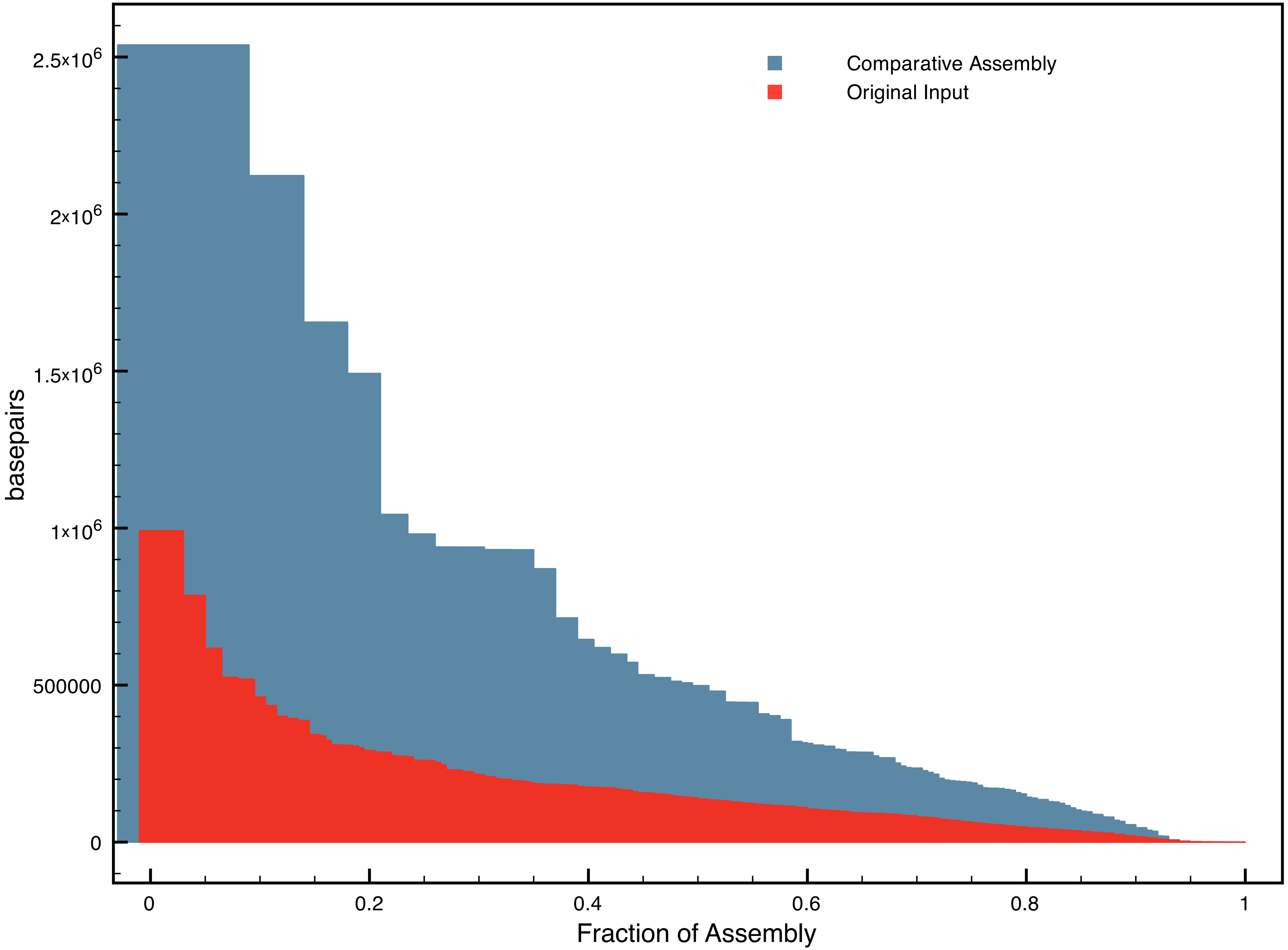
Genome annotation

7,913 genes out of 9,629 have 1:1 Ortholog with *N.crassa*



http://gb2.fungalgenomes.org/gb2/gbrowse/sordaria_macrospora/

Comparative assembly - improve using Synteny



N50=138 kb

N50=498 kb

Brave new world of population genomics

- Resequencing whole genomes quite possible when there are reference genomes available

- For a 40 Mb fungus can easily get SNP results from 1 lane of 36 bp or 72bp sequence with high quality

  - For example in Bd we find about 48k SNPS among strains with 1 lane Illumina ~$1000 - 30Mb genome