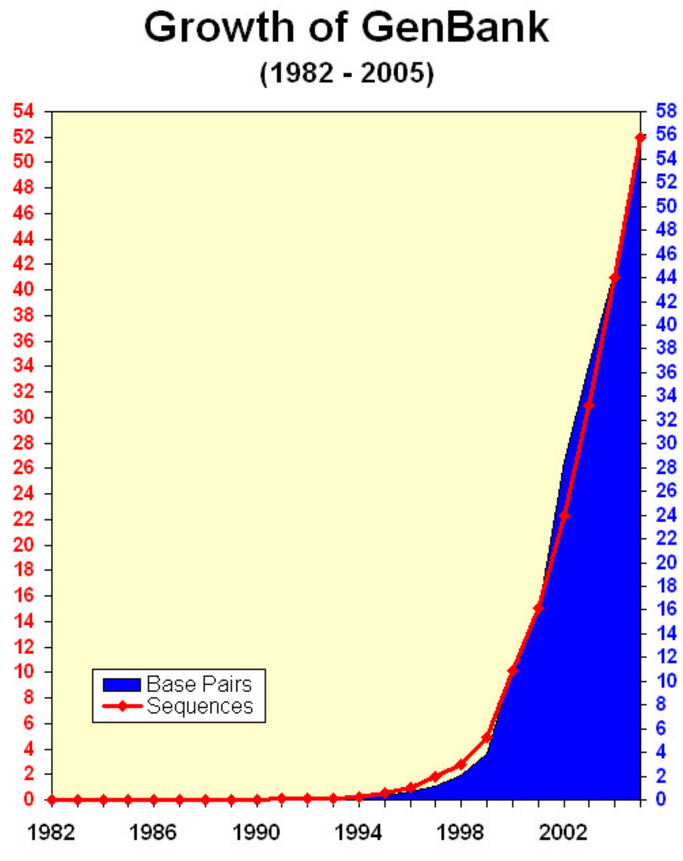
Computer tools used for complex data queries of sequence databases

Jason Stajich
Plant and Microbial Biology
UC Berkeley

Sequence Databases

- Sequence Databases
  - Repository-Archive
    - Annotation of information
- Value is only useful if can extract data from it!
- GenBank-EMBL-DDBJ

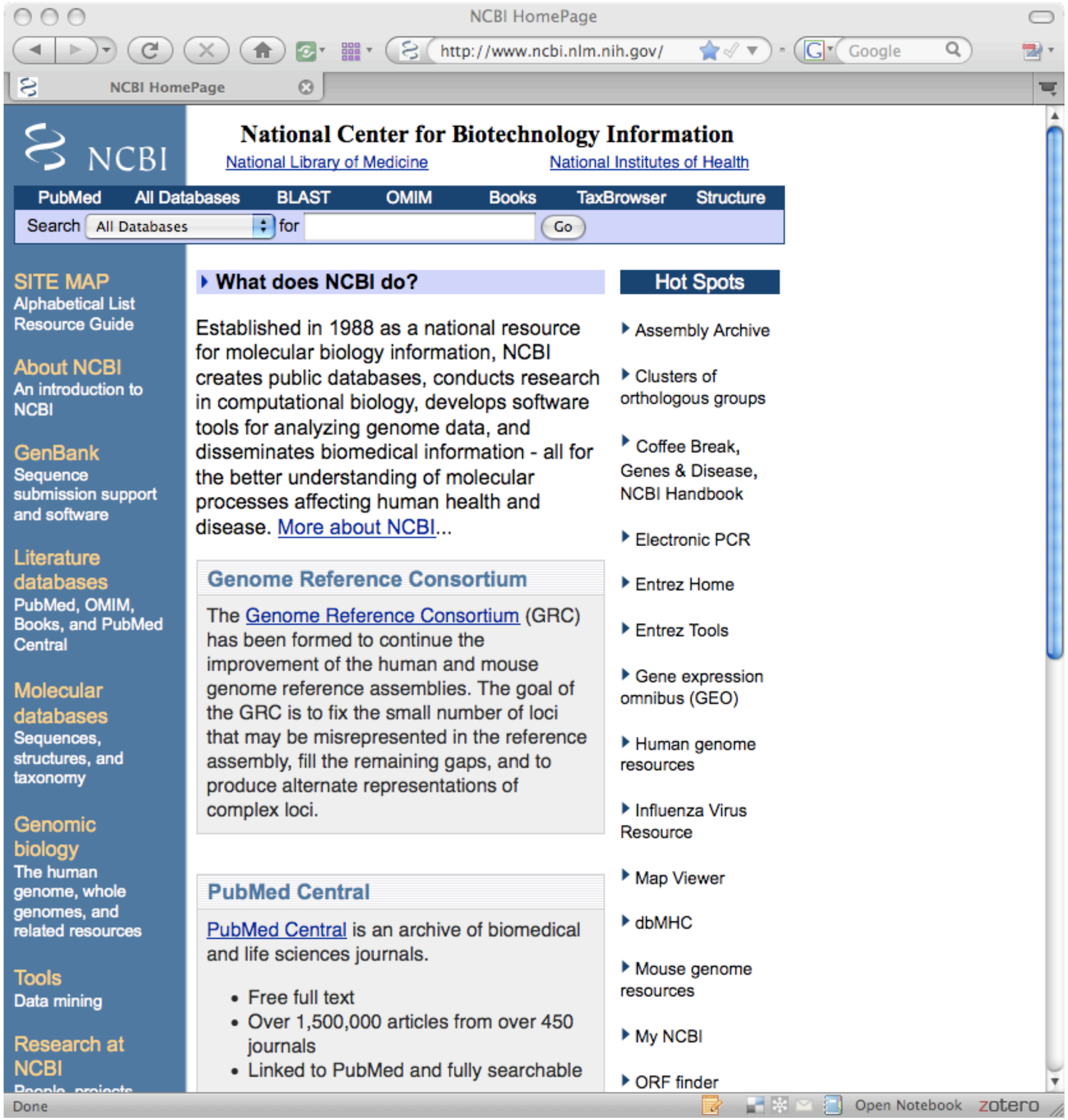


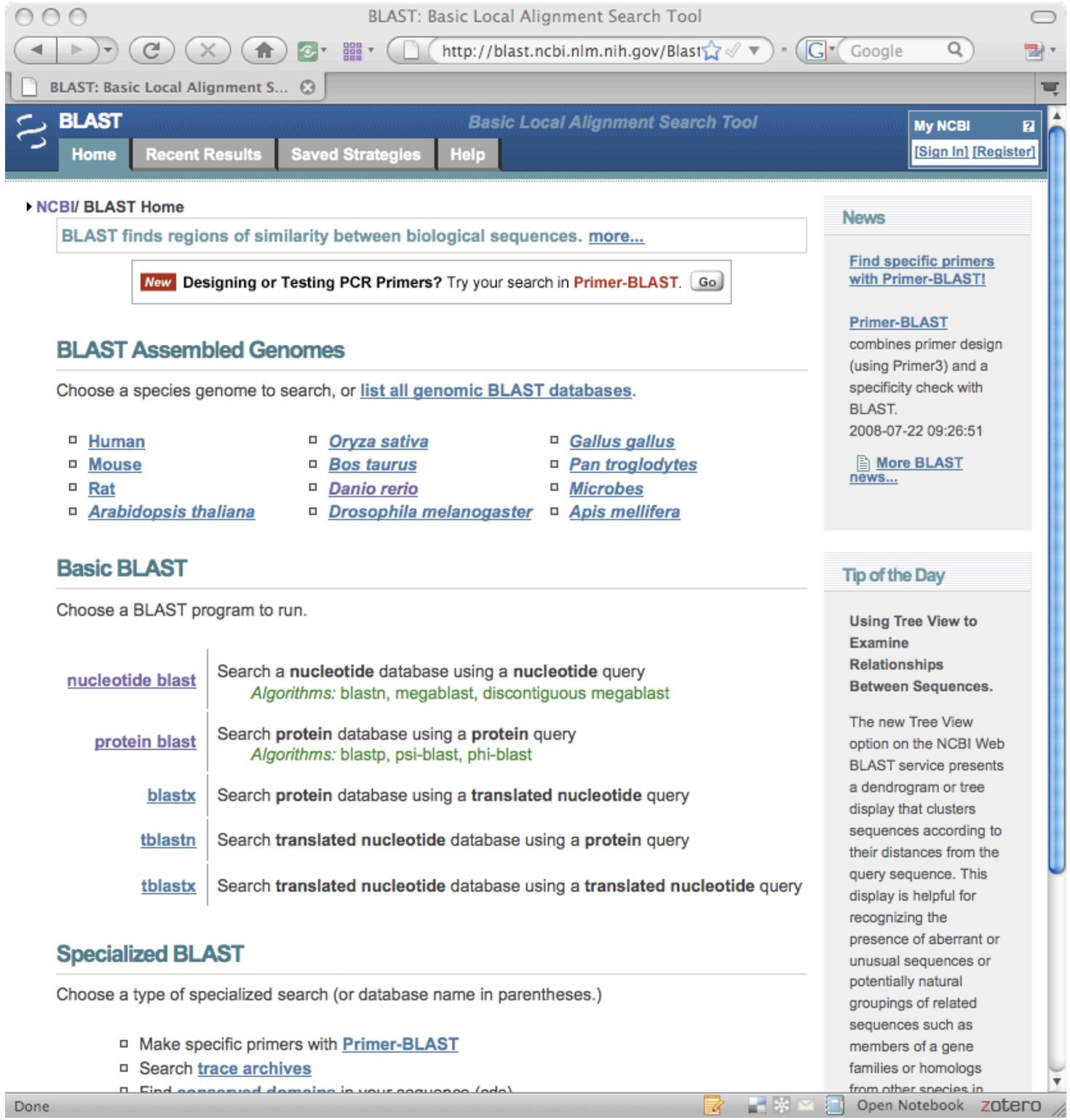
Tool Introduction

- NCBI GenBank + BLAST
  - Identify a sequence
    - Pipelines for automation
- Mining Taxonomy Database
- Programming tools and resources

Individual Example

- For a few sequences from ITS primer amplified from soil library clones, lake water, TOENAILS, etc
- Detailed analysis to determine sequence origin, species identification, etc


NCBI HomePage
http://www.ncbi.nlm.nih.gov/
Google
NCBI HomePage
NCBI
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
PubMed
All Databases
BLAST
OMIM
Books
TaxBrowser
Structure
Search
All Databases
for
Go
SITE MAP
Alphabetical List
Resource Guide
About NCBI
An introduction to NCBI
GenBank
Sequence submission support and software
Literature databases
PubMed, OMIM, Books, and PubMed Central
Molecular databases
Sequences, structures, and taxonomy
Genomic biology
The human genome, whole genomes, and related resources
Tools
Data mining
Research at NCBI
What does NCBI do?
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More about NCBI...
Genome Reference Consortium
The Genome Reference Consortium (GRC) has been formed to continue the improvement of the human and mouse genome reference assemblies. The goal of the GRC is to fix the small number of loci that may be misrepresented in the reference assembly, fill the remaining gaps, and to produce alternate representations of complex loci.
PubMed Central
PubMed Central is an archive of biomedical and life sciences journals.
Free full text
Over 1,500,000 articles from over 450 journals
Linked to PubMed and fully searchable
Hot Spots
Assembly Archive
Clusters of orthologous groups
Coffee Break, Genes & Disease, NCBI Handbook
Electronic PCR
Entrez Home
Entrez Tools
Gene expression omnibus (GEO)
Human genome resources
Influenza Virus Resource
Map Viewer
dbMHC
Mouse genome resources
My NCBI
ORF finder
Done
Open Notebook
zotero


BLAST: Basic Local Alignment Search Tool
http://blast.ncbi.nlm.nih.gov/Blast
Google
BLAST: Basic Local Alignment S...
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST Home
BLAST finds regions of similarity between biological sequences. more...
New Designing or Testing PCR Primers? Try your search in Primer-BLAST. Go
BLAST Assembled Genomes
Choose a species genome to search, or list all genomic BLAST databases.
Human
Mouse
Rat
Arabidopsis thaliana
Oryza sativa
Bos taurus
Danio rerio
Drosophila melanogaster
Gallus gallus
Pan troglodytes
Microbes
Apis mellifera
Basic BLAST
Choose a BLAST program to run.
nucleotide blast
Search a nucleotide database using a nucleotide query
Algorithms: blastn, megablast, discontiguous megablast
protein blast
Search protein database using a protein query
Algorithms: blastp, psi-blast, phi-blast
blastx
Search protein database using a translated nucleotide query
tblastn
Search translated nucleotide database using a protein query
tblastx
Search translated nucleotide database using a translated nucleotide query
Specialized BLAST
Choose a type of specialized search (or database name in parentheses.)
Make specific primers with Primer-BLAST
Search trace archives
News
Find specific primers with Primer-BLAST!
Primer-BLAST combines primer design (using Primer3) and a specificity check with BLAST.
2008-07-22 09:26:51
More BLAST news...
Tip of the Day
Using Tree View to Examine Relationships Between Sequences.
The new Tree View option on the NCBI Web BLAST service presents a dendrogram or tree display that clusters sequences according to their distances from the query sequence. This display is helpful for recognizing the presence of aberrant or unusual sequences or potentially natural groupings of related sequences such as members of a gene families or homologs
Done
Open Notebook
zotero

Nucleotide BLAST: Search nucleotide databases using a nucleotide query

http://blast.ncbi.nlm.nih.gov/Blast

Google

Nucleotide BLAST: Search nucle...

BLAST

Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastn suite: BLASTN programs search nucleotide databases using a nucleotide query. more... Rese

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

Query subrange

From

To

Or, upload file Browse...

Job Title

Enter a descriptive title for your BLAST search

Choose Search Set

Database Human genomic + transcript | Mouse genomic + transcript | Others (nr etc.):

Human genomic plus transcript (Human G+T)

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST

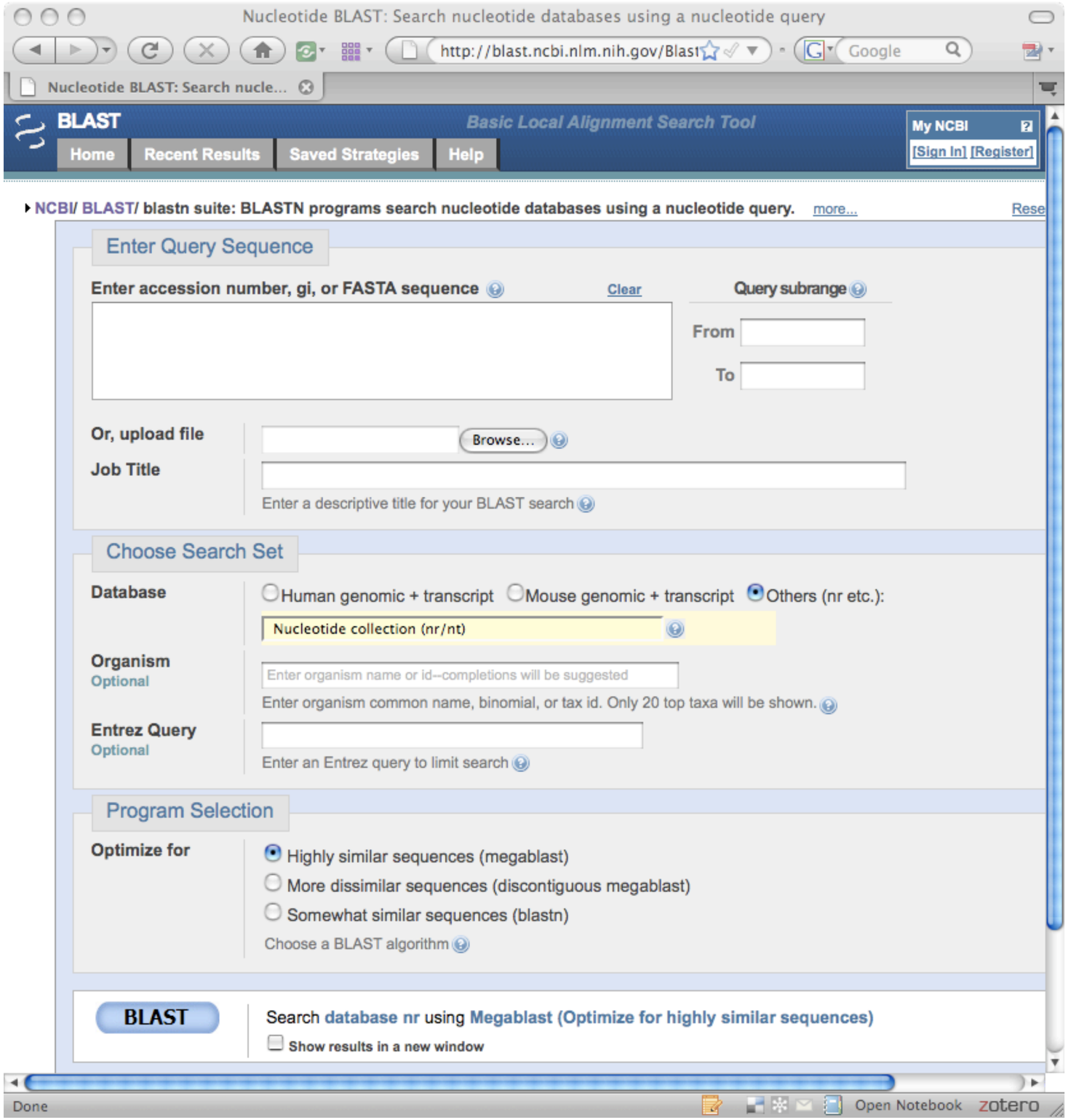
Search database Human G+T using Megablast (Optimize for highly similar sequences)

Show results in a new window

Algorithm parameters

Done

Open Notebook zotero


Nucleotide BLAST: Search nucleotide databases using a nucleotide query
http://blast.ncbi.nlm.nih.gov/Blast
Google
Nucleotide BLAST: Search nucle...
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastn suite: BLASTN programs search nucleotide databases using a nucleotide query.
more...
Rese
Enter Query Sequence
Enter accession number, gi, or FASTA sequence
Clear
Query subrange
From
To
Or, upload file
Browse...
Job Title
Enter a descriptive title for your BLAST search
Choose Search Set
Database
Human genomic + transcript
Mouse genomic + transcript
Others (nr etc.):
Nucleotide collection (nr/nt)
Organism
Optional
Enter organism name or id--completions will be suggested
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Entrez Query
Optional
Enter an Entrez query to limit search
Program Selection
Optimize for
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm
BLAST
Search database nr using Megablast (Optimize for highly similar sequences)
Show results in a new window
Done
Open Notebook
zotero

Nucleotide BLAST: Search nucleotide databases using a nucleotide query

http://blast.ncbi.nlm.nih.gov/Blast

Google

Nucleotide BLAST: Search nucle...

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastn suite: BLASTN programs search nucleotide databases using a nucleotide query. more... Rese

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

Query subrange

From

To

Or, upload file Browse...

Job Title

Enter a descriptive title for your BLAST search

Choose Search Set

Database: Human genomic + transcript | Mouse genomic + transcript | Others (nr etc.):

Nucleotide collection (nr/nt)

Genomic plus Transcript
- Human genomic plus transcript (Human G+T)
- Mouse genomic plus transcript (Mouse G+T)

Other Databases
- Nucleotide collection (nr/nt)
- Reference mRNA sequences (refseq_rna)
- Reference genomic sequences (refseq_genomic)
- NCBI Genomes (chromosome)
- Expressed sequence tags (est)
- Non-human, non-mouse ESTs (est_others)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu_repeats)
- Sequence tagged sites (dbsts)
- Whole-genome shotgun reads (wgs)
- Environmental samples (env_nt)

Organism Optional

top taxa will be shown.

Entrez Query Optional

Program Selectio

Optimize for

ast)

BLAST

Search database nr using Megablast (Optimize for highly similar sequences)

Show results in a new window

Nucleotide BLAST: Search nucleotide databases using a nucleotide query

http://blast.ncbi.nlm.nih.gov/Blast

Google

Nucleotide BLAST: Search nucle...

BLAST

Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI

[Sign In] [Register]

NCBI/ BLAST/ blastn suite: BLASTN programs search nucleotide databases using a nucleotide query. more... Rese

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

```
>P-1-1
GAATAAGTGACCGGCCCTCGGGCCGGGATGTTCACAACCCTTTGTTGTCC
GACTCTGTTGCCTCCGGGGCGACCCTGCCTCCGGGCGGGGGCCCCGGGTG
GACATTTCAAACTCTTGCGTAACTTTGCAGTCTGAGTAAATTTAATTAAT
AAATTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAA
CGCAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATC
```

Query subrange

From

To

Or, upload file Browse...

Job Title P-1-1

Enter a descriptive title for your BLAST search

Choose Search Set

Database Human genomic + transcript | Mouse genomic + transcript | Others (nr etc.):

Nucleotide collection (nr/nt)

Organism Optional: Fungi

- fungi (taxid:4751)
- Fungi/Metazoa group (taxid:33154)
- Fungi (taxid:4751)
- sac fungi (taxid:4890)
- jelly fungi (taxid:5234)
- Fungi/Metazoa incertae sedis (taxid:42461)
- bracket fungi (taxid:5317)
- rust fungi (taxid:5258)
- unclassified Fungi (taxid:89443)
- Funglina (taxid:123758)

p taxa will be shown.

Entrez Query Optional

Program Selectio

Optimize for

st)

Choose a BLAST algorithm

BLAST

Search database nr using Blastn (Optimize for somewhat similar sequences)

Show results in a new window

Choose a BLAST algorithm

BLAST | Search database nr using Megablast (Optimize for highly similar sequences)

Show results in a new window

▼ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 16

Scoring Parameters

Match/Mismatch Scores: 1,-1

Gap Costs: Existence: 5 Extension: 2

Filters and Masking

Filter: Low complexity regions
Species-specific repeats for: Human

Mask: Mask for lookup table only
Mask lower case letters

BLAST | Search database nr using Megablast (Optimize for highly similar sequences)

Show results in a new window

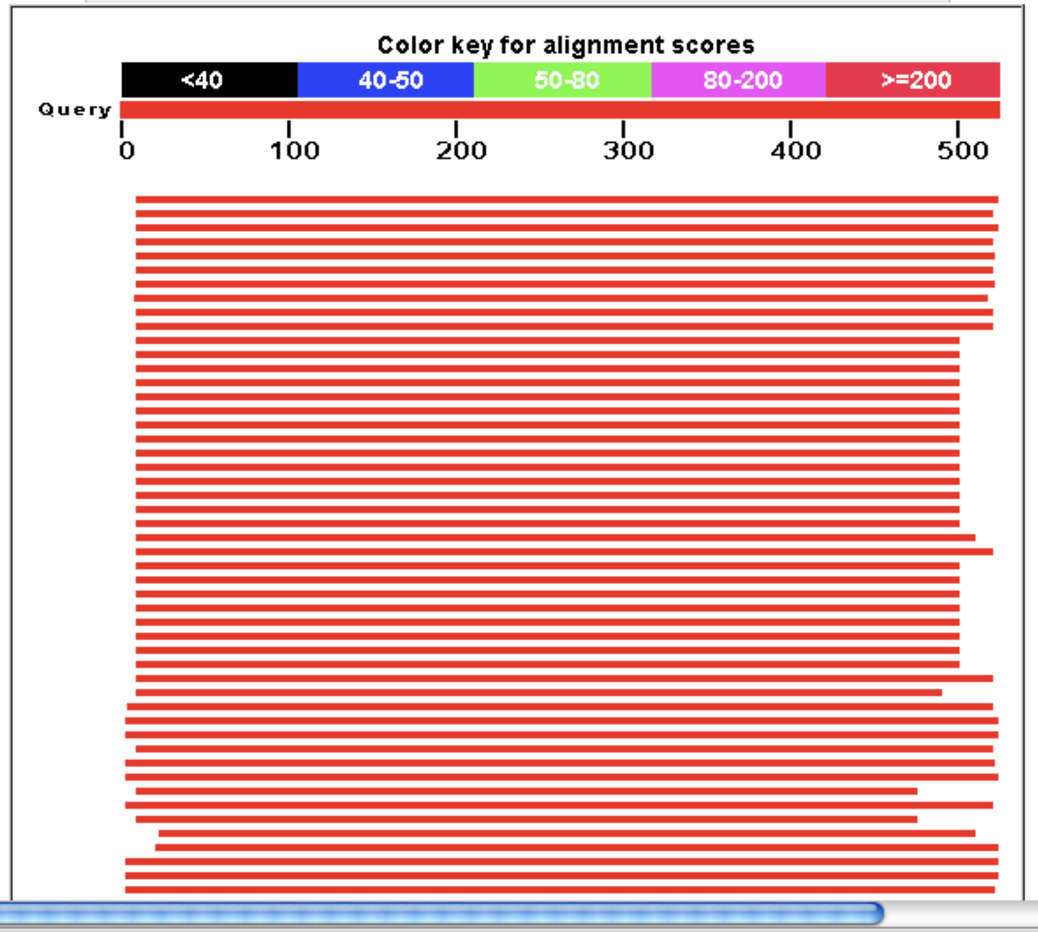
Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
7,129,271 sequences; 24,422,883,224 total letters

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**
Taxonomy reports

Query= P-1-1
Length=526

Distribution of 100 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments



Distance tree of results NEW

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description
EU167574.1	Cladosporium sp. CBS 280.49 small subunit ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence
EU759978.1	Cladosporium sphaerospermum strain IFM 56396 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
EF568046.1	Cladosporium sphaerospermum strain WM 05.10 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene
EF577236.1	Cladosporium cladosporioides 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
AF035674.1	Lacazia loboi internal transcribed spacer 1, 5.8S ribosomal RNA, and internal transcribed spacer 2, partial sequence
DQ875012.1	Cladosporium sp. GP4 18S ribosomal RNA gene, partial sequence
AJ971409.1	Cladosporium sp. MA 4762 partial 18S rRNA gene, ITS1, 5.8S rRNA gene and ITS2, strain MA 4762
L25433.1	Cladosporium sphaerospermum ribosomal RNA gene fragment
DQ279846.1	Uncultured fungus clone G13 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
DQ875017.1	Ascomycete sp. GA4 18S ribosomal RNA gene, partial sequence
EU139855.1	Cladosporium sp. M2077 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
DQ682570.1	Cladosporium sp. IBL 03146 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
AY361982.1	Davidiella tassiana strain ATCC 26362 small subunit ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RI
AY361968.1	Cladosporium cladosporioides strain ATCC 64726 small subunit ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ri
AY625063.1	Cladosporium sphaerospermum strain UAMH 7686 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA ger
AM182174.1	Cladosporium sphaerospermum 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2 and 28S rRNA gene (partial), isolate 2080
AM182171.1	Cladosporium sphaerospermum 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2 and 28S rRNA gene (partial), isolate 1246
AM182168.1	Cladosporium sphaerospermum 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2 and 28S rRNA gene (partial), isolate 1089
AM176749.1	Cladosporium sphaerospermum 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2 and 26S rRNA gene (partial), isolate 1088
AM176686.1	Gliocladium sp. JS1304 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2 and 28S rRNA gene (partial), isolate 1304
AM176685.1	Cladosporium sphaerospermum 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2 and 28S rRNA gene (partial), isolate 2184
AM176684.1	Hyalodendron sp. JS1079 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2 and 28S rRNA gene (partial), isolate 1079
AM176736.1	Hyalodendron sp. JS1248 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2 and 28S rRNA gene (partial), isolate 1248
AM176735.1	Cladosporium sphaerospermum 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2 and 28S rRNA gene (partial), isolate 2105
EF105367.1	Cladosporium sp. CMSJ-2006a 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
AJ971408.1	Cladosporium sp. MA 4764 partial 18S rRNA gene, ITS1, 5.8S rRNA gene and ITS2, strain MA 4764
AM176716.1	Hyalodendron sp. JS1244 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2 and 28S rRNA gene (partial), isolate 1244
AM176715.1	Cladosporium sphaerospermum 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2 and 28S rRNA gene (partial), isolate 1277
AM176714.1	Paecilomyces sp. JS1017 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2 and 28S rRNA gene (partial), isolate 1017
EF568045.1	Cladosporium cladosporioides strain WM 05.11 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, p
AM176721.1	Hyalodendron sp. JS1252 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2 and 28S rRNA gene (partial), isolate 1252
AM176719.1	Cladosporium sphaerospermum 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2 and 28S rRNA gene (partial), isolate 1254
AM902077.1	Uncultured ascomycete ITS region including 18S rRNA gene, ITS1, 5.8S rRNA gene, ITS2 and 28S rRNA gene, clone BF-OTU364
EU825632.1	Uncultured fungus clone yang-Bi83 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequ
DQ092512.1	Cladosporium sp. HKA30 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
AY208804.1	Cladosporium sp. Pr6 internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence
DQ092517.1	Cladosporium sphaerospermum 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequenc
DQ780351.2	Cladosporium sphaerospermum strain CBS 102045 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA ge

EU520106.1	Pleurotus ostreatus isolate NW429 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial seque
EU520178.1	Pleurotus ostreatus isolate NW430 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial seque

Alignments

Get selected sequences | Select all | Deselect all | Distance tree of results

>gb|EU167574.1| Cladosporium sp. CBS 280.49 small subunit ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence
Length=3043

Score = 749 bits (508), Expect = 0.0
Identities = 512/516 (99%), Gaps = 0/516 (0%)
Strand=Plus/Plus

```
Query  11    CCGGCCCTCGGGCCGGGATGTTCACAACCCTTTGTTGTCCGACTCTGTTGCCTCCGGGGC  70
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1703  CCGGCCCTCGGGCCGGGATGTTCACAACCCTTTGTTGTCCGACTCTGTTGCCTCCGGGGC  1762

Query  71    GACCCTGCCTCCGGGCGGGGGCCCCGGGTGGACATTTCAAACTCTTGCGTAACTTTGCAG  130
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1763  GACCCTGCCTCCGGGCGGGGGCCCCGGGTGGACATTTCAAACTCTTGCGTAACTTTGCAG  1822

Query  131   TCTGAGTAAATTTAATTAATAAATTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCAT  190
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1823  TCTGAGTAAATTTAATTAATAAATTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCAT  1882

Query  191   CGATGAAGAACGCAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATC  250
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1883  CGATGAAGAACGCAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATC  1942

Query  251   GAATCTTTGAACGCACATTGCGCCCCCTGGTATTCCGGGGGGCATGCCTGTTCGAGCGTC  310
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1943  GAATCTTTGAACGCACATTGCGCCCCCTGGTATTCCGGGGGGCATGCCTGTTCGAGCGTC  2002

Query  311   ATTTCACCACTCAAGCCTCGCTTGGTATTGGGCGACGCGGTCCGCCGCGCGCCTCAAATC  370
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2003  ATTTCACCACTCAAGCCTCGCTTGGTATTGGGCGACGCGGTCCGCCGCGCGCCTCAAATC  2062

Query  371   GACCGGCTGGGTCTTTCGTCCCCTCAGCGTTGTGGAAACTATTCGCTAAAGGGTGCCGCG  430
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2063  GACCGGCTGGGTCTTTCGTCCCCTCAGCGTTGTGGAAACTATTCGCTAAAGGGTGCCGCG  2122

Query  431   GGAGGCCACGCCGTAAAACAACCCCATTTCTAAGGTTGACCTCGGATCAGGTAGGGATAC  490
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2123  GGAGGCCACGCCGTAAAACAACCCCATTTCTAAGGTTGACCTCGGATCAGGTAGGGATAC  2182

Query  491   CCGCTGAACTTAAGCATATCAAAAGGGGGAAGAAAA  526
             |||||||||||||||||||| | | ||| |||||
Sbjct  2183  CCGCTGAACTTAAGCATATCAATAAGCGGAGGAAAA  2218
```

>gb|EU759978.1| Cladosporium sphaerospermum strain IFM 56396 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete sequence; and 28S ribosomal

1: EF568046. Reports Cladosporium spha...[gi:148535520] Links

Features Sequence

```
LOCUS       EF568046                541 bp    DNA     linear   PLN 21-JUL-2008
DEFINITION  Cladosporium sphaerospermum strain WM 05.10 18S ribosomal RNA gene,
            partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA
            gene, and internal transcribed spacer 2, complete sequence; and 28S
            ribosomal RNA gene, partial sequence.
ACCESSION   EF568046
VERSION     EF568046.1  GI:148535520
KEYWORDS    .
SOURCE      Cladosporium sphaerospermum
  ORGANISM  Cladosporium sphaerospermum
            Eukaryota; Fungi; Dikarya; Ascomycota; Pezizomycotina;
            Dothideomycetes; Dothideomycetidae; Capnodiales; Davidiellaceae;
            mitosporic Davidiellaceae; Cladosporium.
REFERENCE   1  (bases 1 to 541)
  AUTHORS   Kong,F., Tsui,K.M., van de Wiele,N., Chen,S., Sorrell,T., Sun,Y.,
            Huynh,M., Lee,O.C., Halliday,C., Zeng,X., Tong,Z., Chen,X.,
            Porter,G., Robert,V. and Meyer,W.
  TITLE     Establishment of a quality controlled internal transcribed spacer
            region sequence database as basis for routine clinical
            identification of medically relevant fungal pathogens
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 541)
  AUTHORS   Kong,F., Tsui,K.M., van de Wiele,N., Chen,S., Sorrell,T., Sun,Y.,
            Huynh,M., Lee,O.C., Halliday,C., Zeng,X., Tong,Z., Chen,X.,
            Porter,G., Robert,V. and Meyer,W.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-APR-2007) CIDM, ICPMR, Westmead Hospital, Darcy Road,
            Sydney, NSW 2145, Australia
FEATURES             Location/Qualifiers
     source          1..541
                     /organism="Cladosporium sphaerospermum"
                     /mol_type="genomic DNA"
                     /strain="WM 05.10"
                     /db_xref="taxon:92950"
     rRNA            <1..8
                     /product="18S ribosomal RNA"
     misc_RNA        9..165
```

NCBI Taxonomy Browser

Entrez | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search for [ ] as [complete name] lock [Go] [Clear]

Display [3] levels using filter: [none]

- [ ] Nucleotide
- [ ] Nucleotide Core
- [ ] Nucleotide EST
- [ ] Nucleotide GSS
- [ ] Protein
- [ ] Structure
- [ ] Genome Sequences
- [ ] Genome Projects
- [ ] Popset
- [ ] SNP
- [ ] 3D Domains
- [ ] Domains
- [ ] GEO Datasets
- [ ] GEO Profiles
- [ ] UniGene
- [ ] UniSTS
- [ ] PubMed Central
- [ ] Gene
- [ ] HomoloGene
- [ ] MapView
- [ ] LinkOut
- [ ] BLAST
- [ ] TRACE
- [ ] Taxonomy

Lineage (full): root; cellular organisms; Eukaryota; Fungi/Metazoa group; Fungi; Dikarya; Ascomycota; Pezizomycotina; Dothideomycetes; Dothideomycetidae; Capnodiales; Davidiellaceae; mitosporic Davidiellaceae; Cladosporium

---

- **Cladosporium sphaerospermum** *Click on organism name to get more information.*
  - **Cladosporium sp. CBS 117728**

Disclaimer: The NCBI taxonomy database is not an authoritative source for nomenclature or classification - please consult the relevant scientific literature for the most reliable information.

Comments and questions to info@ncbi.nlm.nih.gov

Credits: Joe Bischoff, Mikhail Domrachev, Scott Federhen, Carol Hotton, Detlef Leipe, Vladimir Soussov, Richard Sternberg, Sean Turner.

[Help] [Search] [NLM NIH] [Disclaimer]

Cladosporium sp. CBS 117728

Taxonomy ID: 538541
Rank: varietas
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 4 (Mold Mitochondrial; Protozoan Mitochondrial; Coelenterate Mitochondrial; Mycoplasma; Spiroplasma)

Lineage(full)
cellular organisms; Eukaryota; Fungi/Metazoa group; Fungi; Dikarya; Ascomycota; Pezizomycotina; Dothideomycetes; Dothideomycetidae; Capnodiales; Davidiellaceae; mitosporic Davidiellaceae; Cladosporium; Cladosporium sphaerospermum

Entrez records	
Database name	Direct links
Nucleotide	2
Nucleotide Core	2
Protein	2
Popset	2
Taxonomy	1

Comments and References:

Dugan FM et al. 2008 (unpubl.)
Dugan,F.M., Braun,U., Groenewald,J.Z. and Crous,P.W. A new variety of Cladosporium sphaerospermum from a North American patent collection.

Cladosporium sphaerospermum

Taxonomy ID: 92950
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 4 (Mold Mitochondrial; Protozoan Mitochondrial; Coelenterate Mitochondrial; Mycoplasma; Spiroplasma)
Other names:
synonym: **Cladosporium sphaerospermum Penz.**
includes: **Cladosporium sp. HKB21**

Entrez records

Database name	Subtree links	Direct links
Nucleotide	104	102
Nucleotide Core	104	102
Protein	38	36
Popset	19	19
PubMed Central	24	24
Taxonomy	2	1

Lineage(full)
cellular organisms; Eukaryota; Fungi/Metazoa group; Fungi; Dikarya; Ascomycota; Pezizomycotina; Dothideomycetes; Dothideomycetidae; Capnodiales; Davidiellaceae; mitosporic Davidiellaceae; Cladosporium

External Information Resources (NCBI LinkOut)

LinkOut	Subject	LinkOut Provider
Cladosporium sphaerospermum Penz. 1882	taxonomy/phylogenetic	Global Biodiversity Information Facility
Cladosporium sphaerospermum Penz. 1882	taxonomy/phylogenetic	Index Fungorum
Cladosporium sphaerospermum Penzig	taxonomy/phylogenetic	MycoBank
Cladosporium sphaerospermum	taxonomy/phylogenetic	Systematic Mycology and Microbiology Laboratory, Fungal Databases
Cladosporium sphaerospermum	taxonomy/phylogenetic	TreeBase

Notes:
Groups interested in participating in the LinkOut program should visit the LinkOut home page.
A list of our current non-bibliographic LinkOut providers can be found here.
To see LinkOut links in this lineage click here

Information from sequence entries

Show organism modifiers

NCBI Taxonomy Browser

Search for [ ] as [complete name] [x] lock [Go] [Clear]

Display [3] levels using filter: [none]

[ ] Nucleotide [ ] Nucleotide Core [ ] Nucleotide EST [ ] Nucleotide GSS [ ] Protein [ ] Structure [ ] Genome Sequences
[ ] Genome Projects [ ] Popset [ ] SNP [ ] 3D Domains [ ] Domains [ ] GEO Datasets [ ] GEO Profiles
[ ] UniGene [ ] UniSTS [ ] PubMed Central [ ] Gene [ ] HomoloGene [ ] MapView [ ] LinkOut
[ ] BLAST [ ] TRACE [ ] Taxonomy

Lineage (full): root; cellular organisms; Eukaryota; Fungi/Metazoa group; Fungi; Dikarya; Ascomycota; Pezizomycotina; Dothideomycetes; Dothideomycetidae; Capnodiales; Davidiellaceae; mitosporic Davidiellaceae

---

- **Cladosporium** *Click on organism name to get more information.*
  - **Cladosporium adianticola**
  - **Cladosporium allii-porri**
  - **Cladosporium antarcticum**
  - **Cladosporium breviramosum**
  - **Cladosporium bruhnei**
  - **Cladosporium cladosporioide**
  - **Cladosporium cladosporioides**
    - **Cladosporium cladosporioides f. pisicola**
  - **Cladosporium aff. cladosporioides CBS 109082**
  - **Cladosporium aff. cladosporioides CBS 673.69**
  - **Cladosporium aff. cladosporioides CPC 11606**
  - **Cladosporium aff. cladosporioides CPC 11609**
  - **Cladosporium cf. cladosporioides MU-10**
  - **Cladosporium colocasiae**
  - **Cladosporium coralloides**
  - **Cladosporium cucumerinum**
  - **Cladosporium effusum**
  - **Cladosporium elatum**
  - **Cladosporium funiculosum**
  - **Cladosporium gossypiicola**
  - **Cladosporium herbaroides**
  - **Cladosporium langeronii**
  - **Cladosporium laxicapitulatum**
  - **Cladosporium lignicola**
  - **Cladosporium macrocarpum**

NCBI Taxonomy Browser

Cladosporium

Taxonomy ID: 5498
Rank: genus
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 4 (Mold Mitochondrial; Protozoan Mitochondrial; Coelenterate Mitochondrial; Mycoplasma; Spiroplasma)

Lineage(full)

cellular organisms; Eukaryota; Fungi/Metazoa group; Fungi; Dikarya; Ascomycota; Pezizomycotina; Dothideomycetes; Dothideomycetidae; Capnodiales; Davidiellaceae; mitosporic Davidiellaceae

Entrez records

Database name	Subtree links	Direct links
Nucleotide	1,026	-
Nucleotide Core	1,026	-
Protein	388	-
Popset	79	76
PubMed Central	307	257
Taxonomy	202	1

External Information Resources (NCBI LinkOut)

LinkOut	Subject	LinkOut Provider
Cladosporium	taxonomy/phylogenetic	Global Biodiversity Information Facility
Cladosporium _	taxonomy/phylogenetic	
Cladosporium	taxonomy/phylogenetic	Index Fungorum
search ING	taxonomy/phylogenetic	Index Nominum Genericorum
Cladosporium Link Ex Fries, 1815	taxonomy/phylogenetic	Integrated Taxonomic Information System
Cladosporium Link	taxonomy/phylogenetic	MycoBank
Cladosporium	taxonomy/phylogenetic	TreeBase

Notes:

Groups interested in participating in the LinkOut program should visit the LinkOut home page.
A list of our current non-bibliographic LinkOut providers can be found here.
To see LinkOut links in this lineage click here

Information from sequence entries

Show organism modifiers

Found 1026 nucleotide sequences. Nucleotide [1026]

1: EU167592 Reports Links

Cladosporium sp. CBS 266.53 small subunit ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence
gi|162415172|gb|EU167592.1|[162415172]

2: EU167586 Reports Links

Cladosporium sp. CBS 282.49 small subunit ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence
gi|162415166|gb|EU167586.1|[162415166]

3: EU167574 Reports Links

Cladosporium sp. CBS 280.49 small subunit ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence
gi|162415154|gb|EU167574.1|[162415154]

4: AY787691 Reports Links

Cladosporium magnusianum 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
gi|194306792|gb|AY787691.2|[194306792]

5: FM179794 Reports Links

Cladosporium sp. SPRY16 partial 18S rRNA gene, ITS1, 5.8S rRNA gene, ITS2 and partial 28S rRNA gene, clone SPRY16
gi|194305721|emb|FM179794.1|[194305721]

NCBI

Entrez Nucleotide

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search Nucleotide for txid5498[Organism:exp] AND ("internal transcribed spacer" Preview Go Clear

Limits | Preview/Index | History | Clipboard | Details

About Entrez

Entrez Nucleotide
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

My NCBI (Cubby)

Related resources
BLAST

Reference sequence project

Search for Genes

Submit to GenBank

Search for full length cDNAs

- Enter terms and click Preview to see only the number of search results.
- To save search indefinitely, click query # and select Save in My NCBI.
- To combine searches use #search, e.g., #2 AND #3 or click query # for more options.

Search	Most Recent Queries	Time	Result
#8	Search **txid5498[Organism:exp] AND ("internal transcribed spacer" OR ITS)**	14:45:27	425
#5	Search **txid5498[Organism:exp]**	14:44:25	1026

Add Term(s) to Query or View Index:

- Enter a term in the text box; use the pull-down menu to specify a search field.
- Click Preview to add terms to the query box and see the number of search results, or click Index to view terms within a field.

All Fields Preview Index

Click AND OR NOT to add a term to the query box

NCBI

Entrez Nucleotide

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search Nucleotide for (Cladosporium[Organism]) Go Clear

Limits | Preview/Index | History | Clipboard | Details

About Entrez
Entrez Nucleotide
Help | FAQ
Entrez Tools
Check sequence revision history
LinkOut
My NCBI (Cubby)
Related resources
BLAST
Reference sequence project
Search for Genes
Submit to GenBank
Search for full length cDNAs

Limits: **Genomic DNA/RNA, Genomic DNA/RNA, published in the last 30 days**

- Use All Fields pull-down menu to specify a field.
- If search fields tags are used enclose in square brackets, e.g., rubella [ti].
- More help on using limits is available here.

Limited to:

Fields: All Fields

Exclude: ☐ STSs ☐ working draft ☐ TPA ☐ patents

Molecule: Genomic DNA/RNA

Gene Location: Genomic DNA/RNA

Segmented Sequences: Any

Only from: Any

Published in the last: 30 days

Modified in the last: Any Date

(Cladosporium[Organism]) – Nucleotide Results

http://www.ncbi.nlm.nih.gov/sites/entrez

ITS RDB fungal

(Cladosporium[Organism]) – N... | NCBI Sequence Viewer v2.0 | NCBI Sequence Viewer v2.0

NCBI

Entrez Nucleotide

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search Nucleotide for (Cladosporium[Organism]) Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Limits: **Genomic DNA/RNA, Genomic DNA/RNA, published in the last 30 days**

Found 6 nucleotide sequences. Nucleotide [6]

Display Summary Show 20 Sort by Send to

All: 6 | Bacteria: 0 | RefSeq: 0 | mRNA: 0

Items 1 - 6 of 6

One page.

1: FM179794 Reports Links
Cladosporium sp. SPRY16 partial 18S rRNA gene, ITS1, 5.8S rRNA gene, ITS2 and partial 28S rRNA gene, clone SPRY16
gi|194305721|emb|FM179794.1|[194305721]

2: FM179793 Reports Links
Cladosporium sp. SPRY17 partial 5.8S rRNA gene, ITS2 and partial 28S rRNA gene, clone SPRY17
gi|194305720|emb|FM179793.1|[194305720]

3: EU729719 Reports Links
Cladosporium sp. NIOCC F13 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence
gi|194173352|gb|EU729719.1|[194173352]

4: EU729712 Reports Links
Cladosporium cladosporioides strain NIOCC F8 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence
gi|194173345|gb|EU729712.1|[194173345]

5: EU729711 Reports Links
Cladosporium cladosporioides strain NIOCC F5 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
gi|194173344|gb|EU729711.1|[194173344]

Top Organisms [Tree]
- Cladosporium cladosporioides *(2)*
- Cladosporium sp. NIOCC F13 *(1)*
- Cladosporium sp. LY 4.2 *(1)*
- Cladosporium sp. SPRY17 *(1)*
- Cladosporium sp. SPRY16 *(1)*

Done

Open Notebook zotero

My NCBI

http://www.ncbi.nlm.nih.gov/entrez/cubby.fcgi?call=0oS8EgTd0C11BKvkZeDdUm0pKJ-uym%3BQuer

Save Search

Your search in Nucleotide

```
(Cladosporium[Organism]) AND (biomol_genomic[PROP] AND gene_in_genomic[PROP] AND "last
30 days"[PDAT])
```

Enter a name for your search: (Cladosporium) (biomol_genomic gene_in_genomic "last 30

Would you like to receive e-mail updates of new search results? (•) No ( ) Yes

OK Cancel

Done Open Notebook zotero

High throughput sequencing

Technique	Year Launched	Read Length	Reads per Run	Throughput per run	Cost per GB of seq
ABI 3730	2000	800-1100	96	0.1Mb	>2500K
454 (Roche) GSFLX	2004	250-400	400K	~100 Mb	$80-100K
Solexa (Illumina)	2006	35-50	~100M	5 GB	$4K
ABI SOLiD	2007	25-35	~150M	5-6GB	$4K
Helicos Heliscope	2008	28-30	85M	2GB	?

No Amplification Needed

6 months=1 Terabase

Press Releases: 2nd July 2008

Fifteen human genomes each week

The Wellcome Trust Sanger Institute Hits 1 Terabase

The Wellcome Trust Sanger Institute has sequenced the equivalent of 300 human genomes in just over six months. The Institute has just reached the staggering total of 1,000,000,000,000 letters of genetic code that will be read by researchers worldwide, helping them to understand the role of genes in health and disease. Scientists will be able to answer questions unthinkable even a few years ago and human medical genetics will be transformed.

The amount of data is remarkable: every two minutes, the Institute produces as much sequence as was deposited in the first five years of the international DNA sequence databases, which started in 1982. It is a global milestone.

"I am delighted that our rapid adoption of next-generation sequencing technologies has been so successful in driving forward our biomedical research," says Dr Harold Swerdlow, Head of Sequencing Technology at the Wellcome Trust Sanger Institute. *"Our internal projects, our work with external collaborators and our participation in major international programmes are all benefiting from our success."*

> “ Our internal projects, our work with external collaborators and our participation in major international programmes are all benefiting from our success. ”
>
> **Dr Harold Swerdlow**

The Institute has major roles in projects such as The 1000 Genomes Project, The International Cancer Genome Consortium and the second round of the Wellcome Trust Case Control Consortium, all of which will depend on DNA sequence to uncover genetics variants that are important for human disease. Next-generation sequencing is also enabling the Institute's own research portfolio.

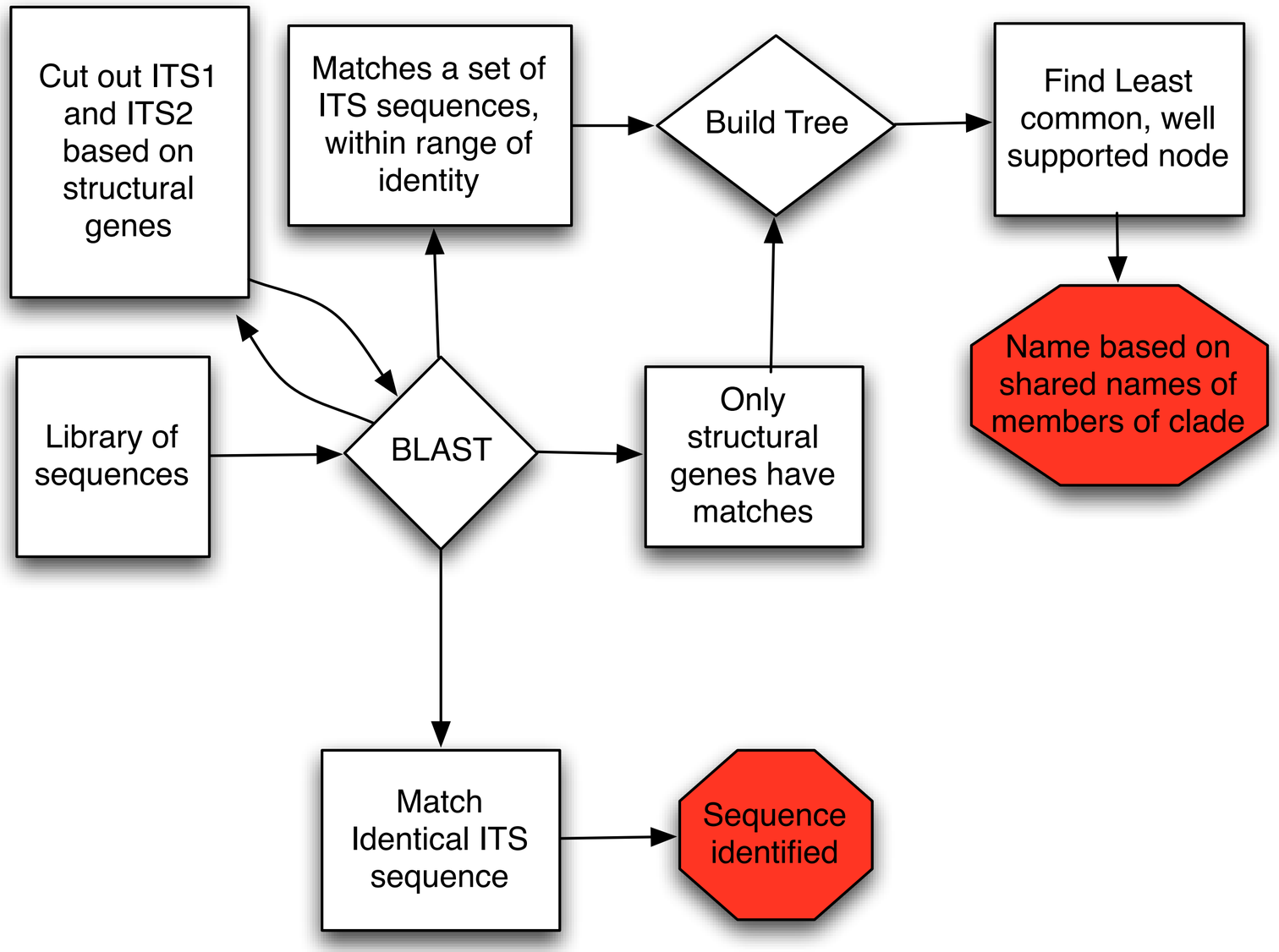
"The Sanger Institute is positioned to take on challenges and to answer questions that are daunting to most," says Professor Allan Bradley, Director. *"We can explore important biomedical questions in a way that few can match, and next-generation sequencing is a vital part of that quest."*

If printed in 12 point Courier, the DNA sequenced by the Institute would stretch round the world 63 times.

What to do with many more sequences

- What to do with thousands of reads?
- Pipelines for analyses
- Amplification of just a single molecule (barcode)
- Metagenomic sampling without amplification?

Identify a species for unknown sequence



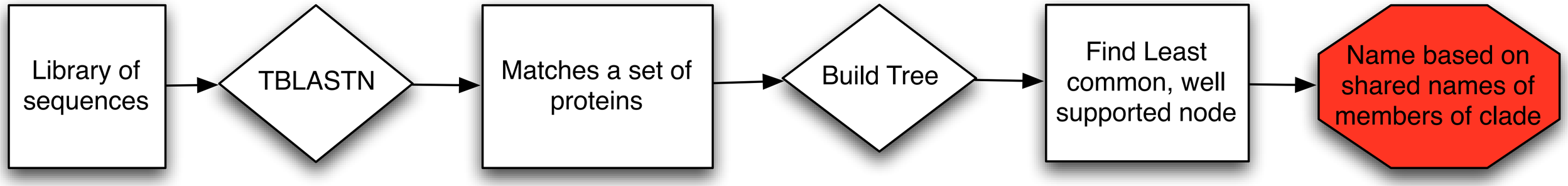
Strategies

- If a barcode sequence doesn't match exactly to anything what to do?
  - Build trees from structural genes
    - Sequence another molecule (AFTOL genes, others).
- Obtain tree and name based on well supported clade from conserved molecules
- Other methods for training

Metagenomic approaches

- Sample sequences from environment
- Cannot link sequences together, each read is independent sample
- Pros
  - Without primer amplification
  - Next Gen sequencing removes need to clone
- Problems
  - Random sampling of sequences, how to link together?

Phylogenomic Binning



Assess community structure comparing the clades
Walk up Taxonomy or species tree and count representatives per clade

Get selected sequences | Select all | Deselect all | Distance tree of results

>gb|EU167574.1| Cladosporium sp. CBS 280.49 small subunit ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence
Length=3043

Score = 749 bits (508), Expect = 0.0
Identities = 512/516 (99%), Gaps = 0/516 (0%)
Strand=Plus/Plus

```
Query  11    CCGGCCCTCGGGCCGGGATGTTCACAACCCTTTGTTGTCCGACTCTGTTGCCTCCGGGGC  70
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1703  CCGGCCCTCGGGCCGGGATGTTCACAACCCTTTGTTGTCCGACTCTGTTGCCTCCGGGGC  1762

Query  71    GACCCTGCCTCCGGGCGGGGGCCCCGGGTGGACATTTCAAACTCTTGCGTAACTTTGCAG  130
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1763  GACCCTGCCTCCGGGCGGGGGCCCCGGGTGGACATTTCAAACTCTTGCGTAACTTTGCAG  1822

Query  131   TCTGAGTAAATTTAATTAATAAATTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCAT  190
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1823  TCTGAGTAAATTTAATTAATAAATTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCAT  1882

Query  191   CGATGAAGAACGCAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATC  250
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1883  CGATGAAGAACGCAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATC  1942

Query  251   GAATCTTTGAACGCACATTGCGCCCCCTGGTATTCCGGGGGGCATGCCTGTTCGAGCGTC  310
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1943  GAATCTTTGAACGCACATTGCGCCCCCTGGTATTCCGGGGGGCATGCCTGTTCGAGCGTC  2002

Query  311   ATTTCACCACTCAAGCCTCGCTTGGTATTGGGCGACGCGGTCCGCCGCGCGCCTCAAATC  370
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2003  ATTTCACCACTCAAGCCTCGCTTGGTATTGGGCGACGCGGTCCGCCGCGCGCCTCAAATC  2062

Query  371   GACCGGCTGGGTCTTTCGTCCCCTCAGCGTTGTGGAAACTATTCGCTAAAGGGTGCCGCG  430
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2063  GACCGGCTGGGTCTTTCGTCCCCTCAGCGTTGTGGAAACTATTCGCTAAAGGGTGCCGCG  2122

Query  431   GGAGGCCACGCCGTAAAACAACCCCATTTCTAAGGTTGACCTCGGATCAGGTAGGGATAC  490
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2123  GGAGGCCACGCCGTAAAACAACCCCATTTCTAAGGTTGACCTCGGATCAGGTAGGGATAC  2182

Query  491   CCGCTGAACTTAAGCATATCAAAAGGGGGAAGAAAA  526
             ||||||||||||||||||| | | ||| |||||
Sbjct  2183  CCGCTGAACTTAAGCATATCAATAAGCGGAGGAAAA  2218
```

>gb|EU759978.1| Cladosporium sphaerospermum strain IFM 56396 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete sequence; and 28S ribosomal

Done

Using the Taxonomy Database

- Hierarchy of species relationships
  - Can lookup a taxa
    - Query for taxa at different levels of hierarchy
- (May not reflect current accepted relationships or naming)
  Sometimes Anamorph/Teleomorph linkage is not made.
- Every sequence in Genbank is attached to a taxonomy identifier

NCBI

Entrez Taxonomy

About Entrez

Entrez Taxonomy
Help
Linkout tutorial

Taxonomy home

Taxonomy browser

Taxonomy common tree

Taxonomy resources

Taxonomy Statistics

Taxonomy FTP site

Taxonomy FAQs

How to reference the NCBI taxonomy database

How to create LinkOut links from the NCBI taxonomy

The NCBI Entrez Taxonomy Homepage

The NCBI taxonomy database contains the names of all organisms that are represented in the genetic databases with at least one nucleotide or protein sequence. Click on the tree if you want to browse the taxonomic structure or retrieve sequence data for a particular group of organisms.

These are direct links to some of the organisms commonly used in molecular research projects:

- Arabidopsis thaliana
- Bos taurus
- Caenorhabditis elegans
- Chlamydomonas reinhardtii
- Danio rerio (zebrafish)
- Dictyostelium discoideum
- Drosophila melanogaster
- Escherichia coli
- Hepatitis C virus
- Homo sapiens
- Magnaporthe grisea
- Mus musculus
- Mycoplasma pneumoniae
- Neurospora crassa
- Oryza sativa
- Plasmodium falciparum
- Pneumocystis carinii
- Rattus norvegicus
- Saccharomyces cerevisiae
- Schizosaccharomyces pombe
- Takifugu rubripes
- Xenopus laevis
- Zea mays

Display Common Tree

Comments and questions to info@ncbi.nlm.nih.gov
Credits: Joe Bischoff, Mikhail Domrachev, Scott Federhen, Carol Hotton, Detlef Leipe, Vladimir Soussov, Richard Sternberg, Sean Turner.

Main Page

BioPerl

main links
- Main Page
- Getting Started
- Downloads
- Installation
- Recent changes
- Random page

documentation
- Quick Start
- FAQ
- HOWTOs
- BioPerl Tutorial
- Tutorials
- Deobfuscator
- Browse Modules

community
- News
- Mailing lists
- Supporting BioPerl
- About this site

development
- Developer Information
- Advanced BioPerl
- Subversion
- API Docs
- Bugs

search
Go Search

Introduction

This is the BioPerl project's community documentation site. You can read up on Getting BioPerl, Installing BioPerl, and Getting Started using the toolkit. Advanced BioPerl covers a number of topics once you've got the hang of things. Also use the Frequently Asked Questions, HOWTOs, and the BioPerl Tutorials as starting place for learning about the toolkit's components.

More specific information for developers can be found here and on the sidebar.

There is a short History of BioPerl with background on the project and Lincoln Stein's article on How Perl Saved the Human Genome Project.

See what BioPerl Users are using the toolkit for and what publications cite the toolkit.

The toolkit is divided into several packages, most people will only want to deal with the Core package.

- Core package provides the main parsers, this is the basic package and it's required by all the other packages (bioperl-live SVN directory)
- Run package provides wrappers for executing some 60 common bioinformatics applications (bioperl-run in SVN)
- BioPerl db package is a subproject to store sequence and annotation data in a BioSQL relational database (bioperl-db in SVN).
- Network package parses and analyzes protein-protein interaction data (bioperl-network in SVN).
- See all the available packages via Downloads page

API documentation for each module at doc.bioperl.org and each module page on this site is linked to the doc site and the CPAN perldoc documentation. There are lists of the documented Core modules, Run modules, DB modules, and Ext modules available on this site and the external API documentation.

Current Events

General

- We are now gathering comments on the proposed core module changes.
- Bioperl 1.5.2 has been released! Check the Release page for further details.
- Deobfuscator released to help navigate the API documentation better.
- We welcome new developers to help with projects in the project priority list.
- There is list of orphan modules which need a volunteer to maintain them.
- BOSC 2007 was held July 19-20, 2007 in conjunction with ISMB2007 in Vienna, Austria. BOSC 2006 was held on Aug 4-5, 2006 at ISMB 2006 in Fortaleza, Brazil.
- See the News page for latest mailing list traffic and news. Also see the RSS feeds for easy ways to subscribe to mailing list and development feeds.

Developers

- Migration to the new Subversion is complete. Information is now available on the BioPerl migration to Subversion and using Subversion with BioPerl.
- Feature/Annotation changes introduced prior to the 1.5 release have been rolled back in preparation for a new stable release. The Rollback page has the detailed notes for the rollback; currently all tests related to this pass.
- We are planning tentative steps in making GFF3-compatible output for the various BioPerl classes. See the code audit page for details; contribute by discussing suggestions on the mail list and donating code!

Module:Bio::DB::Taxonomy

Pdoc documentation: Bio::DB::Taxonomy

CPAN documentation: Bio::DB::Taxonomy

Example usage

Here is some example usage that shows how to gather children from a node

```
#!/usr/bin/perl -w
use strict;
use Bio::DB::Taxonomy;

my $idx_dir = '/tmp';

my ($nodefile,$namesfile) = ('nodes.dmp','names.dmp');
my $db = new Bio::DB::Taxonomy(-source    => 'flatfile',
                               -nodesfile => $nodefile,
                               -namesfile => $namesfile,
                               -directory => $idx_dir);
my $node = $db->get_Taxonomy_Node(-taxonid => '33090');
print $node->id, " ", $node->scientific_name, " ", $node->rank, "\n";
# to only get children that are of a particular rank in the taxonomy test if their rank is 'species' for example
my @extant_children = grep { $_->rank eq 'species' } $node->get_all_Descendents;

for my $child ( @extant_children ) {
    print "id is ", $child->id, "\n"; # NCBI taxa id
    print "rank is ", $child->rank, "\n"; # e.g. species
    print "scientific name is ", $child->scientific_name, "\n"; # scientific name
}
```

Ammended for Bio::Taxon missing each_Descendent implementation

Note that this was fixed in CVS 18-Jun-2007 so you don't need to add the extra overridden each_Descendent code.

```
#!/usr/bin/perl -w
use strict;
use Bio::DB::Taxonomy;
use Bio::Taxon;

sub Bio::Taxon::each_Descendent {
    my ($self) = shift;
    my $db ||= $self->db_handle || return;
    return $db->each_Descendent($self);
```

BioPerl

- Open-source library of Perl modules for writing scripts for life sciences data - mostly molecular and sequence data.
- Tools for parsing:
  Sequence data (GenBank, FASTA)
  DB Search results (BLAST)
  Alignments (ClustalW)
- Tools for manipulating and parsing phylogenetic trees.
- Interface to Taxonomy and GenBank databases
- Tools for visualizing genomic data (Gbrowse)
- http://bioperl.org

Parse BLAST

Reference: Gish, W. (1996-2000) http://blast.wustl.edu

Query= BOSS_DROME Bride of sevenless protein precursor.
        (896 letters)

Database: wormpep87
           20,881 sequences; 9,238,759 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

Sequences producing High-scoring Segment Pairs:			High Score	Smallest Sum Probability P(N)	N
F35H10.10	CE24945	status:Partially_confirmed TR:Q20073...	182	4.9e-11	1
M02H5.2	CE25951	status:Predicted TR:Q966H5 protein_id:...	86	0.15	1
ZC506.4	CE01682	locus:mgl-1 metatrophic glutamate recept...	91	0.18	1
F23D12.2	CE05700	status:Partially_confirmed TR:Q19761 ...	73	0.45	3

>F35H10.10 CE24945   status:Partially_confirmed TR:Q20073
           protein_id:AAA81683.2
         Length = 1404

 Score = 182 (69.1 bits), Expect = 4.9e-11, P = 4.9e-11
 Identities = 75/315 (23%), Positives = 149/315 (47%)

```
Query:   511 YPFLFDGESVMFWRIKMDTWVATGLTAAILGLIATLAILVFIVVRISLGDVFEGNPTTSI 570
             Y  +F+  +   WR     +V   L   ++  +  +A+LV ++V++ L  V +GN +  I
Sbjct:  1006 YQSVFEHITTGHWRDHPHNYVLLALITVLV--VVAIAVLVLVLVKLYLR-VVKGNQSLGI 1062

Query:   571 LLLLSLILVFCSFVPYSIEYVGEQRNSHVTFEDAQTLNTLCAVRVFIMTLVYCFVFSLLL 630
              LL+ +I++      YS  +          F+     +++C +RV +  L Y   F +++
Sbjct:  1063 SLLIGIIIL------YSTAFF-------FVFDPT---DSVCRLRVILHGLGYTICFGVMI 1106

Query:   631 CRAVMLASIGSEG-GFLSHVNGYIQAVICAFSVVAQVGMSVQLLVVMHVASETVSCENIY 689
              +A  L +  + G G   H++ +   ++  F V  Q +S+   +++  V   N+
Sbjct:  1107 AKATQLRNAETLGFGTAIHISFWNYWLLLFFIVGVQIALSISWFLEPFMSTIGVIDTNVQ 1166
```

Result

Hit

HSP

Reference: Gish, W. (1996-2000) http://blast.wustl.edu

Query= BOSS_DROME Bride of sevenless protein precursor.
        (896 letters)

Database: wormpep87
           20,881 sequences; 9,238,759 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

Result

Sequences producing High-scoring Segment Pairs:			High Score	Smallest Sum Probability P(N)	N
F35H10.10	CE24945	status:Partially_confirmed TR:Q20073...	182	4.9e-11	1
M02H5.2	CE25951	status:Predicted TR:Q966H5 protein_id:...	86	0.15	1
ZC506.4	CE01682	locus:mgl-1 metatrophic glutamate recept...	91	0.18	1
F23D12.2	CE05700	status:Partially_confirmed TR:Q19761 ...	73	0.45	3

Hit

>F35H10.10 CE24945 status:Partially_confirmed TR:Q20073
           protein_id:AAA81683.2
         Length = 1404

 Score = 182 (69.1 bits), Expect = 4.9e-11, P = 4.9e-11
 Identities = 75/315 (23%), Positives = 149/315 (47%)

```
Query:   511 YPFLFDGESVMFWRIKMDTWVATGLTAAILGLIATLAILVFIVVRISLGDVFEGNPTTSI 570
             Y  +F+  +   WR     +V   L   ++  +  +A+LV ++V++ L  V +GN +  I
Sbjct:  1006 YQSVFEHITTGHWRDHPHNYVLLALITVLV--VVAIAVLVLVLVKLYLR-VVKGNQSLGI 1062

Query:   571 LLLLSLILVFCSFVPYSIEYVGEQRNSHVTFEDAQTLNTLCAVRVFIMTLVYCFVFSLLL 630
              LL+ +I++      YS  +          F+     +++C +RV +  L Y   F +++
Sbjct:  1063 SLLIGIIIL------YSTAFF-------FVFDPT---DSVCRLRVILHGLGYTICFGVMI 1106

Query:   631 CRAVMLASIGSEG-GFLSHVNGYIQAVICAFSVVAQVGMSVQLLVVMHVASETVSCENIY 689
              +A  L +  + G G   H++ +  ++  F  V  Q+ +S+   +  +++  V   N+
Sbjct:  1107 AKATQLRNAETLGFGTAIHISFWNYWLLLFFIVGVQIALSISWFLEPFMSTIGVIDTNVQ 1166
```

HSP

```
Reference:  Gish, W. (1996-2000) http://blast.wustl.edu

Query=  BOSS_DROME Bride of sevenless protein precursor.
        (896 letters)

Database:  wormpep87
           20,881 sequences; 9,238,759 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                   Smallest
                                                                     Sum
                                                             High  Probability
Sequences producing High-scoring Segment Pairs:             Score  P(N)      N

F35H10.10 CE24945    status:Partially_confirmed TR:Q20073...  182  4.9e-11   1
M02H5.2   CE25951    status:Predicted TR:Q966H5 protein_id:...  86  0.15      1
ZC506.4   CE01682  locus:mgl-1 metatrophic glutamate recept...  91  0.18      1
F23D12.2  CE05700    status:Partially_confirmed TR:Q19761 ...   73  0.45      3

>F35H10.10 CE24945    status:Partially_confirmed TR:Q20073
           protein_id:AAA81683.2
        Length = 1404

 Score = 182 (69.1 bits), Expect = 4.9e-11, P = 4.9e-11
 Identities = 75/315 (23%), Positives = 149/315 (47%)

Query:   511 YPFLFDGESVMFWRIKMDTWVATGLTAAILGLIATLAILVFIVVRISLGDVFEGNPTTSI 570
             Y  +F+  +   WR     +V   L   ++  +  +A+LV ++V++ L  V +GN +  I
Sbjct:  1006 YQSVFEHITTGHWRDHPHNYVLLALITVLV--VVAIAVLVLVLVKLYLR-VVKGNQSLGI 1062

Query:   571 LLLLSLILVFCSFVPYSIEYVGEQRNSHVTFEDAQTLNTLCAVRVFIMTLVYCFVFSLLL 630
              LL+ +I++       YS  +          F+     +++C +RV +  L Y   F +++
Sbjct:  1063 SLLIGIIIL------YSTAFF-------FVFDPT---DSVCRLRVILHGLGYTICFGVMI 1106

Query:   631 CRAVMLASIGSEG-GFLSHVNGYIQAVICAFSVVAQVGMSVQLLVVMHVASETVSCENIY 689
              +A  L +  + G G   H++ +   ++  F V  Q+ +S+   +   +++  V   N+
Sbjct:  1107 AKATQLRNAETLGFGTAIHISFWNYWLLLFFIVGVQIALSISWFLEPFMSTIGVIDTNVQ 1166
```

Result

Hit

HSP

```
Reference:  Gish, W. (1996-2000) http://blast.wustl.edu

Query=  BOSS_DROME Bride of sevenless protein precursor.
        (896 letters)

Database:  wormpep87
           20,881 sequences; 9,238,759 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                     Smallest
                                                                       Sum
                                                              High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)      N

F35H10.10 CE24945    status:Partially_confirmed TR:Q20073...   182  4.9e-11   1
M02H5.2   CE25951    status:Predicted TR:Q966H5 protein_id:...  86  0.15      1
ZC506.4   CE01682  locus:mgl-1 metatrophic glutamate recept...  91  0.18      1
F23D12.2  CE05700    status:Partially_confirmed TR:Q19761 ...   73  0.45      3

>F35H10.10 CE24945    status:Partially_confirmed TR:Q20073
           protein_id:AAA81683.2
        Length = 1404

 Score = 182 (69.1 bits), Expect = 4.9e-11, P = 4.9e-11
 Identities = 75/315 (23%), Positives = 149/315 (47%)

Query:   511 YPFLFDGESVMFWRIKMDTWVATGLTAAILGLIATLAILVFIVVRISLGDVFEGNPTTSI 570
             Y  +F+  +   WR     +V   L   ++  +  +A+LV ++V++ L  V +GN +  I
Sbjct:  1006 YQSVFEHITTGHWRDHPHNYVLLALITVLV--VVAIAVLVLVLVKLYLR-VVKGNQSLGI 1062

Query:   571 LLLLSLILVFCSFVPYSIEYVGEQRNSHVTFEDAQTLNTLCAVRVFIMTLVYCFVFSLLL 630
              LL+ +I++      YS  +          F+     +++C +RV +  L Y   F +++
Sbjct:  1063 SLLIGIIIL------YSTAFF-------FVFDPT---DSVCRLRVILHGLGYTICFGVMI 1106

Query:   631 CRAVMLASIGSEG-GFLSHVNGYIQAVICAFSVVAQVGMSVQLLVVMHVASETVSCENIY 689
              +A  L +  + G G   H++ +   ++  F V  Q+ +S++   +   +++  V   N+
Sbjct:  1107 AKATQLRNAETLGFGTAIHISFWNYWLLLFFIVGVQIALSISWFLEPFMSTIGVIDTNVQ 1166
```

Result

Hit

HSP

```
use Bio::SearchIO;
my $cutoff = ’0.001’;
my $file = ‘BOSS_Ce.BLASTP’,
my $in = Bio::SearchIO->new(-format => ‘blast’,
                            -file   => $file);
while( my $r = $in->next_result ) {
  print "Query is: ", $r->query_name, " ", $r->query_description," ",
        $r->query_length," aa\n";
  print " Matrix was ", $r->get_parameter(’matrix’), "\n";
  while( my $h = $r->next_hit ) {
    last if $h->significance > $cutoff;
    print "Hit is ", $h->name, "\n";
    while( my $hsp = $h->next_hsp ) {
     print " HSP Len is ", $hsp->length(’total’), " ",
           " E-value is ", $hsp->evalue,
           " Bit score ", $hsp->score, " \n",
           " Query loc: ",$hsp->query->start, " ", $hsp->query->end," ",
           " Sbject loc: ",$hsp->hit->start, " ", $hsp->hit->end,"\n";
    }
  }
}
```

Parsing Result

```
Query is: BOSS_DROME Bride of sevenless protein
precursor. 896 aa
Matrix was BLOSUM62
Hit is F35H10.10
HSP Len is 315 E-value is 4.9e-11 Bit score 182
Query loc: 511 813 Sbject loc: 1006 1298
HSP Len is 28 E-value is 1.4e-09 Bit score 39
Query loc: 508 535 Sbject loc: 427 454
```

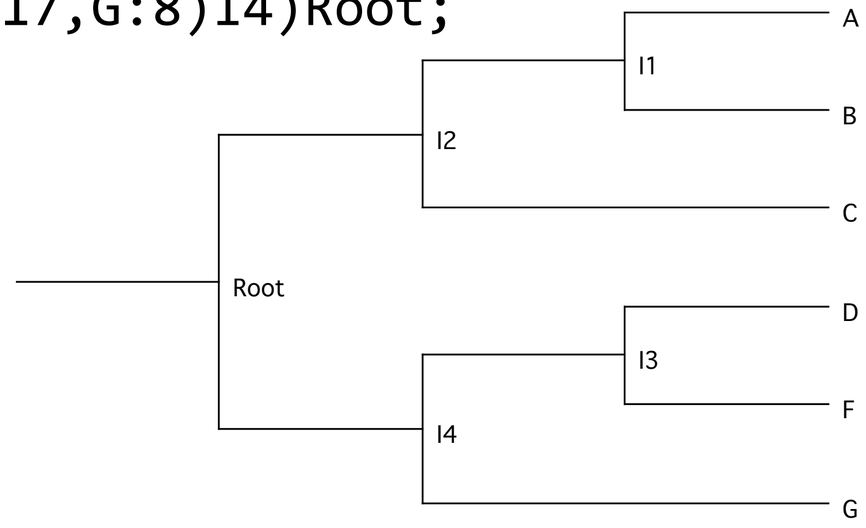
Processing a Tree

Walking up the tree (tips to root)

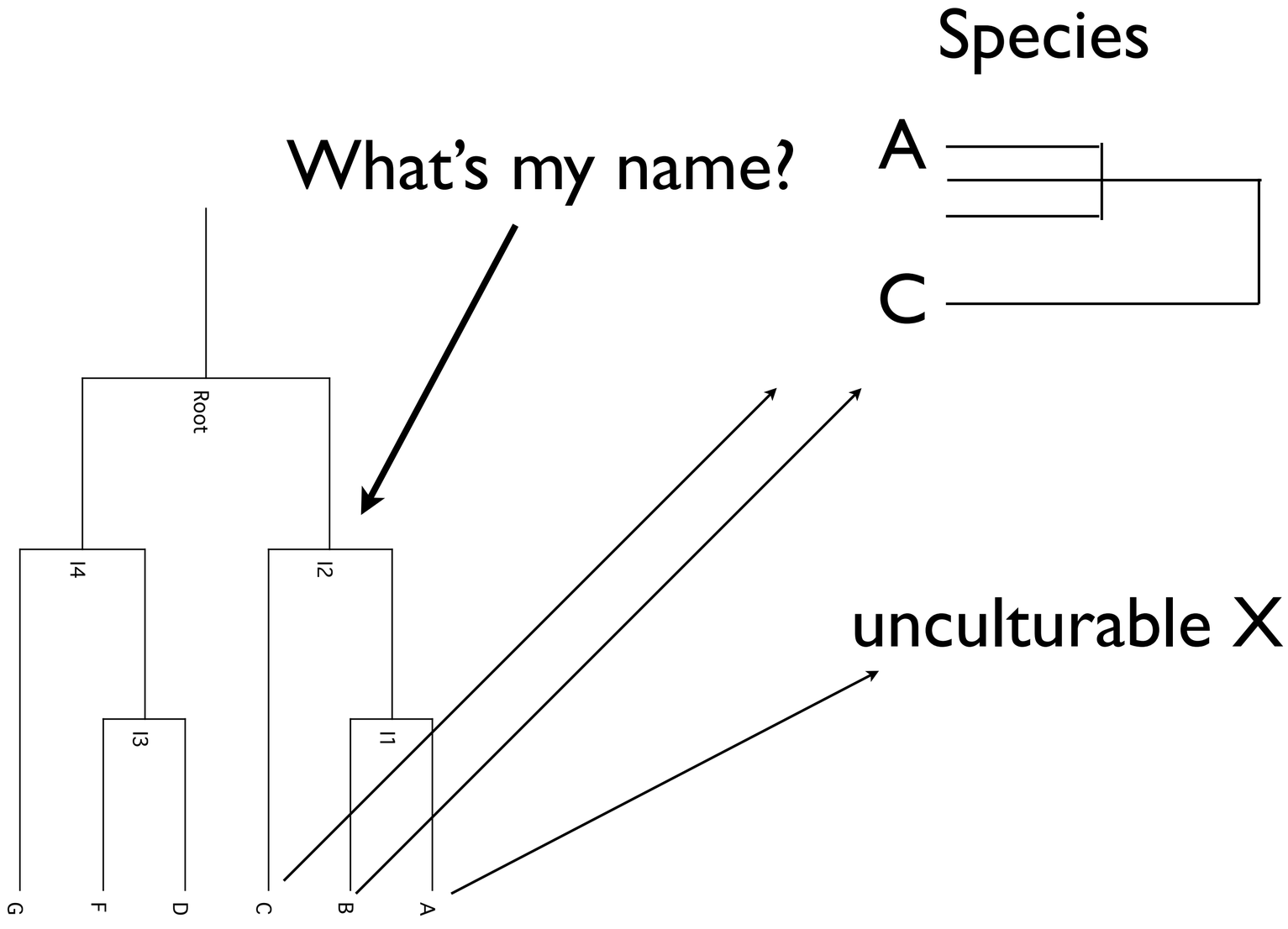
```
if( my $tree = $in->next_tree ) {
   my @tips = grep { $_->is_Leaf } $tree->get_nodes;
   for my $node ( @tips ) {
      my @path;
      while( defined $node) {
        push @path, $node->id;
        $node = $node->ancestor;
      }
      print join(“,“, @path), “\n”;
   }
}
__DATA__
(((A:10,B:11)I1:2,C:5)I2,((D:7,F:6)I3:17,G:8)I4)Root;
```

Output

```
C,I2,Root
A,I1,I2,Root
B,I1,I2,Root
G,I4,Root
D,I3,I4,Root
F,I3,I4,Root
```



Conceptually looking for node name



unculturable X

http://fungalgenomes.org/blog

The Hyphal Tip: Fungal Genomes and Comparative Genomics

Digesting the fungal genomes

Cochliobolus genome released

Posted on July 17th, 2008 by Jason Stajich · No Comments

Just noticed that the JGI has released the *Cochliobolus heterostrophus* genome sequence at their site predicting 9,633 protein-coding genes. Torrey Mesa Research Institute had access to a sequence many years ago, but it isn't until now that public version of this genome is available. *Cochliobolus* is has been a model plant pathogen system and its production of T-Toxin by a PKS gene (Yang et al).

Categories: pezizomycota

→ No Comments

Time lapse Coprinopsis growth

Posted on July 14th, 2008 by Jason Stajich · No Comments

FUNGALGENOMES TWITTER

- fungalgenomes: Agaricus bisporus moving through JGI pipeline: per @mike_challen http://is.gd/UGW July 16, 2008
- fungalgenomes: WP 2.6 upgraded with SVN. Yah projects making repositories available. July 14, 2008
- fungalgenomes: Reloaded Coccidioides Gbrowse for Development server with new tracks and Broad v2.1 annotation. July 12, 2008
- fungalgenomes: Installed WP 2.6-beta3 on Blog site. July 9, 2008
- fungalgenomes: @mike_challen I think it would be interesting exp - maybe would have layers of information attributed to users so they could be filtered. July 8, 2008

July 2008

M	T	W	T	F	S	S
	1	2	3	4	5	6
7	8	9	10	11	12	13

SEARCH IT!

To search, type and hit enter

RECENT ENTRIES

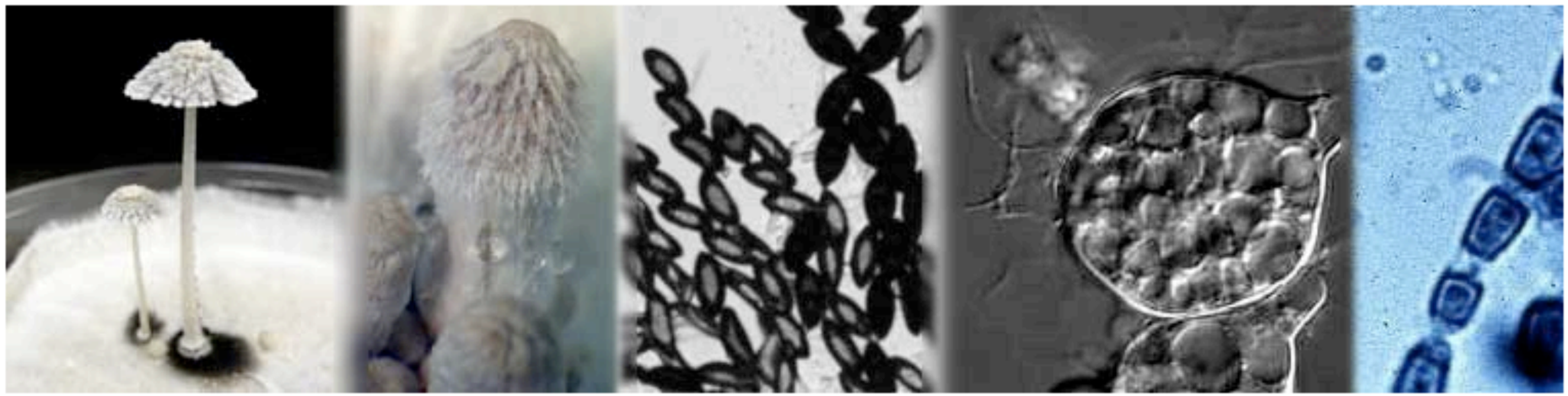
- Cochliobolus genome released 7.17
- Time lapse Coprinopsis growth 7.14
- Basidiomycete Research Networks 7.14
- AAM Releases "The Fungal Kingdom" Report 7.9
- Theobroma cacao to be sequenced, Oompa Loompa genome to follow. 6.26
- Fungal genome assembly from short-read sequences 6.16
- Will you always be able to satisfy that chocolate craving? 6.15
- Amphibian skin bacteria shown to fight off Batrachochytrium dendrobatidis. 6.5
- Penicillium marneffei project 6.3
- Basidiomycete genomes galore 6.1
- Visit the archives for more!

http://fungalgenomes.org/wiki/

Main Page | Projects | Genomes | Recent changes | Fungal Genome Home | Blog | Search

FungalGenome

Fungal Genome Links



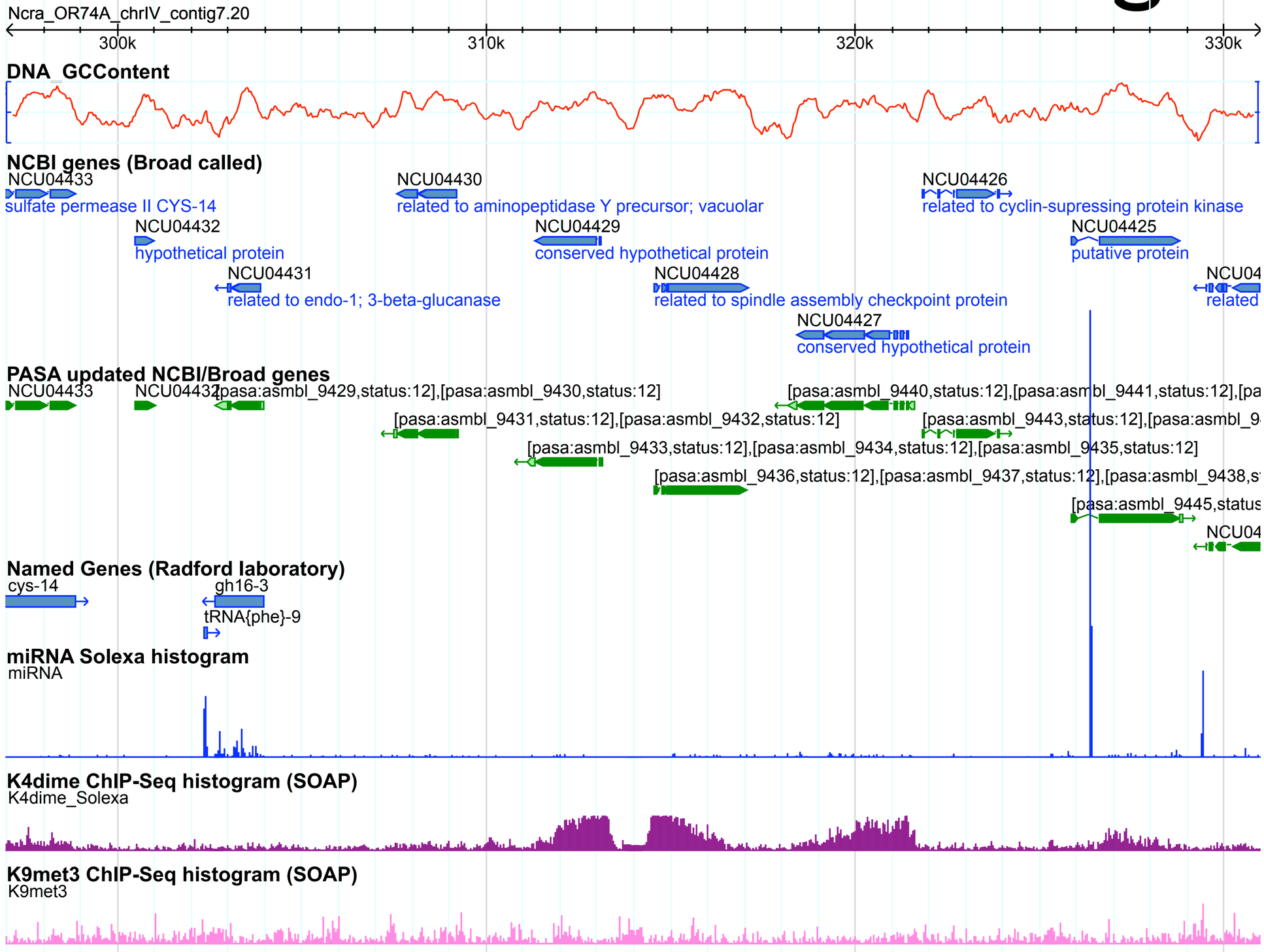
Links and references for the currently available fungal genome sequences or proposed (and austensibly funded) fungal genomes.

Contents [hide]

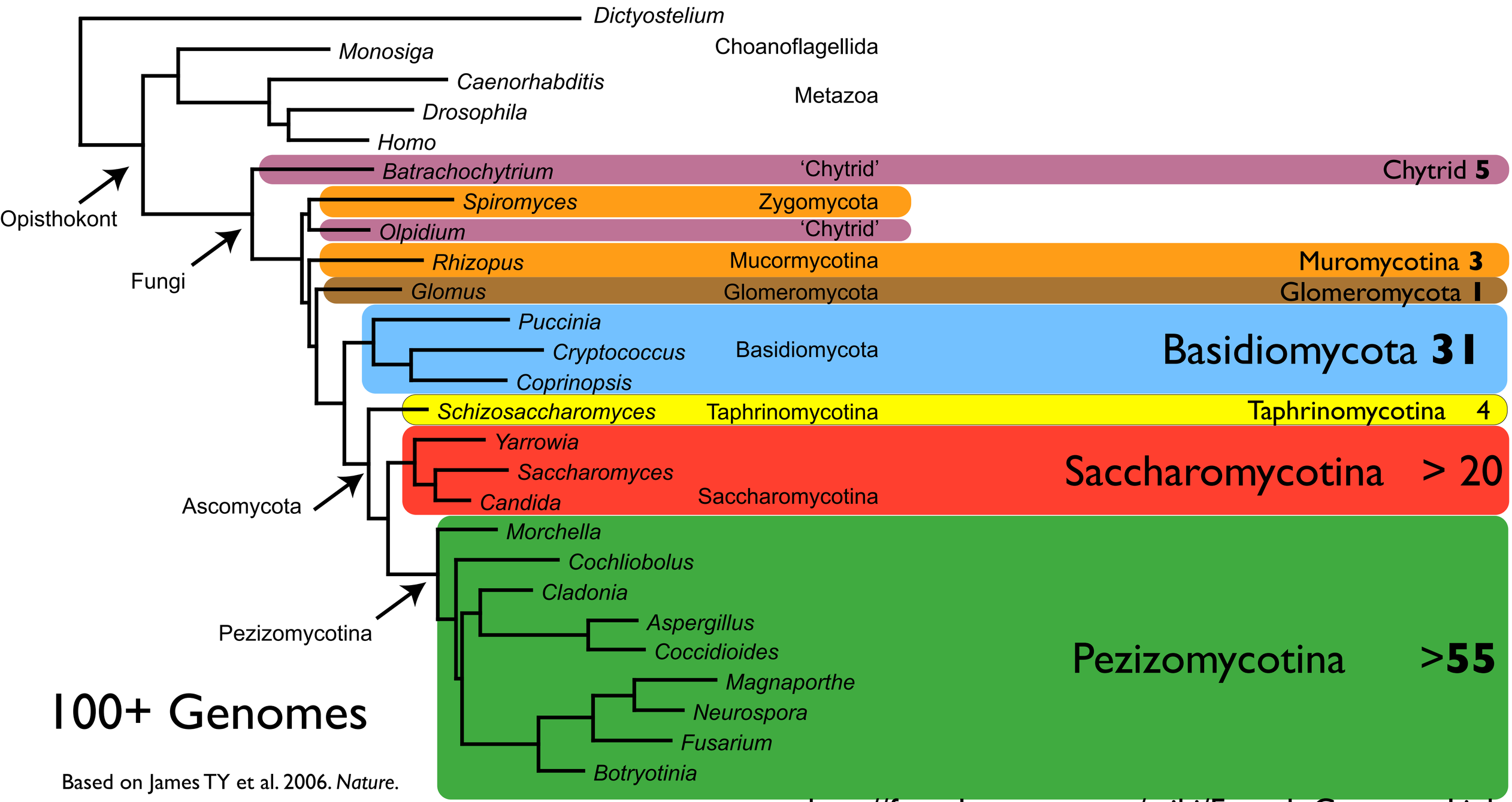
Views

- Article
- Discussion
- Edit
- History
- Protect
- Delete
- Move
- Unwatch

Genome Browser data integration



Genome samples from fungi



100+ Genomes

Based on James TY et al. 2006. *Nature*.

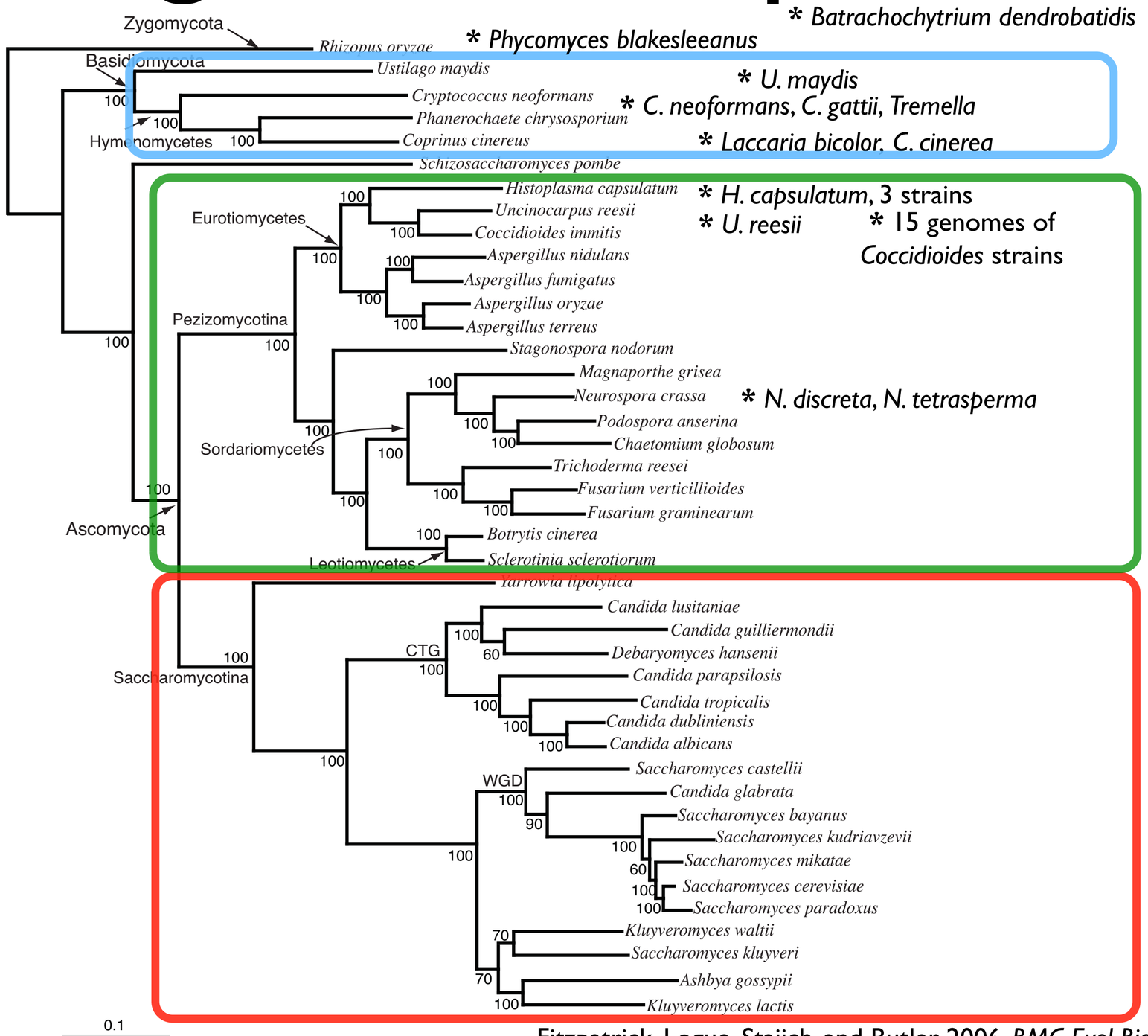
http://fungalgenomes.org/wiki/Fungal_Genome_Links

Phylogeny from genome sequences

- 100s-1000s of gene phylogenies, Concatenated and individual gene trees and consensus

- Generally recapitulate what is found from multi-locus studies of 2-4 genes

- Identify conflicting genes to find cases of Incomplete lineage sorting, bad orthology assignment, or possible horizontal transfer



Fitzpatrick, Logue, Stajich, and Butler 2006. *BMC Evol Biol*

Utility of Genomes?

- Inventory and organization of genes
- Additional marker identification
  - Predict Microsattelites and other highly polymorphic regions for population studies
  - Find more single-copy genes for orthology

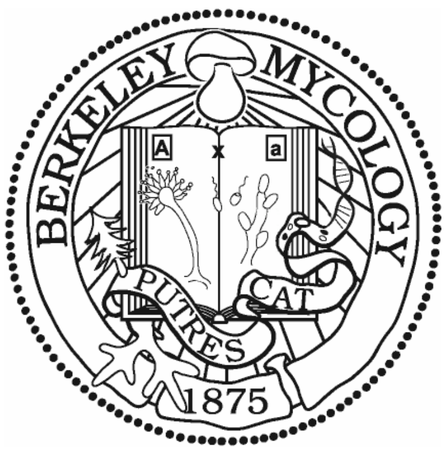
Links

NCBI - http://www.ncbi.nlm.nih.gov

BioPerl - http://bioperl.org

GMOD & Gbrowse - http://www.gmod.org

Fungal Genomes & News Blog - http://fungalgenomes.org/



Miller Institute