

Las bases de datos del GenBank

Francisco Pando



The annual “NAR January Database issue”

<http://www.oxfordjournals.org/nar/database/c/>

OXFORD JOURNALS

Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE ARCHIVE

Oxford Journals > Life Sciences > Nucleic Acids Research > Database Summary Paper

2014 NAR Database Summary Paper

- ▶ [Compilation Paper](#)
- ▶ [Category List](#)
- ▶ [Alphabetical List](#)
- ▶ [Category/Paper List](#)
- ▶ [Search Summary Papers](#)

Nucleotide Sequence Databases

- International Nucleotide Sequence Database Collaboration
 - BioSample
 - DDBJ - DNA Data Bank of Japan
 - EBI patent sequences
 - European Genome-phenome Archive (EGA)
 - European Nucleotide Archive
 - GenBank®
 - NCBI BioSample/BioProject
 - neXtProt
 - The Sequence Read Archive (SRA)
- Coding and non-coding DNA
- Gene structure, introns and exons, splice sites
- Transcriptional regulator sites and transcription factors

RNA sequence databases

Protein sequence databases

Structure Databases

Genomics Databases (non-vertebrate)

Metabolic and Signaling Pathways

Human and other Vertebrate Genomes

Human Genes and Diseases

Microarray Data and other Gene Expression Databases

Proteomics Resources

Other Molecular Biology Databases

Organelle databases

Plant databases

Immunological databases

Cell biology

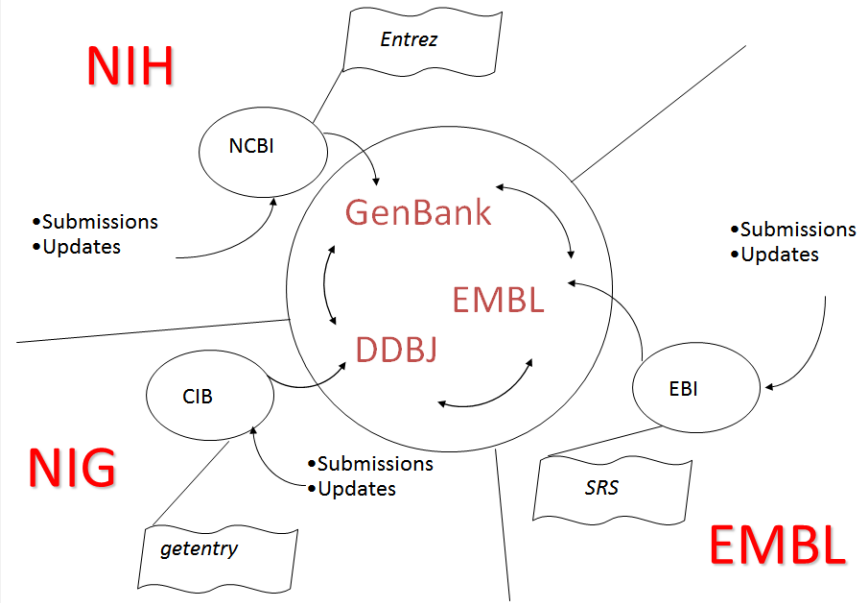
GenBank

GenBank[®] is a comprehensive database that contains publicly available nucleotide sequences for more than 300 000 organisms named at the genus level or lower, obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects. Accession numbers are assigned by GenBank staff upon receipt. Daily data exchange with the European Molecular Biology Laboratory Nucleotide Sequence Database in Europe and the DNA Data Bank of Japan ensures worldwide coverage. GenBank is accessible through the NCBI Entrez retrieval system, which integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed. To access GenBank and its related retrieval and analysis services, begin at the NCBI Homepage using the link provided.

<http://www.ncbi.nlm.nih.gov/>

<http://www.ncbi.nlm.nih.gov/books/NBK21105/>

<http://www.oxfordjournals.org/nar/database/summary/3>

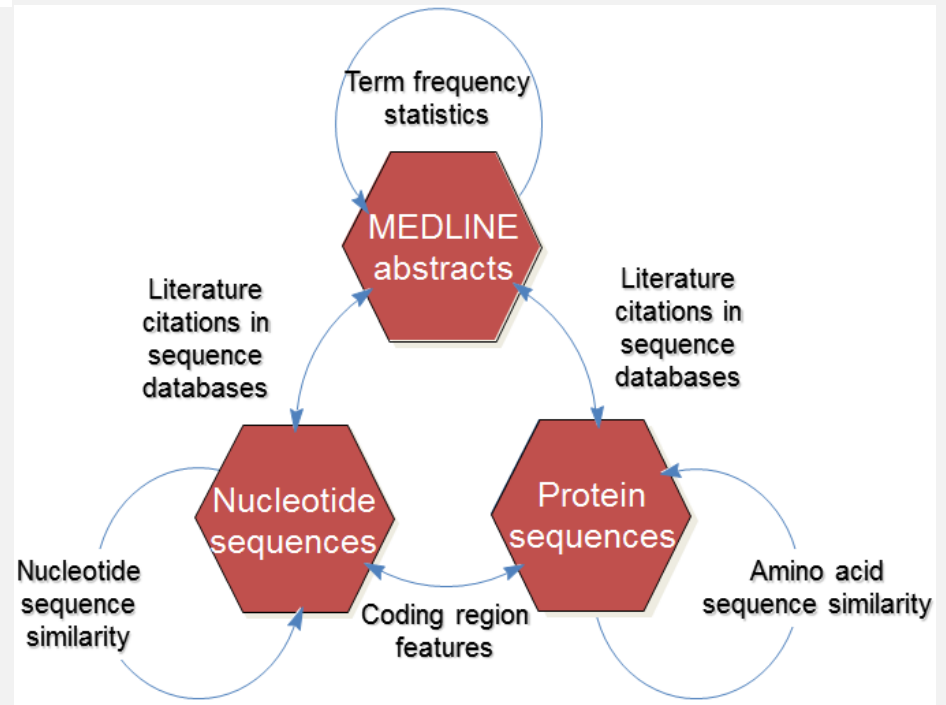


If part of a GenBank nucleotide sequence encodes a protein, a conceptual translation – called a coding region or coding sequence (CDS) – is annotated. A protein accession number (a "protein id") is assigned to the translation product and is noted on the GenBank record.

This protein id is linked to a record for the protein sequence in NCBI's protein databases



NCBI receives and processes about 20,000 direct submission sequences per month, in addition to the approximately 200,000 bulk submissions that are processed automatically



GQuery

NCBI Global Cross-database Search

Search NCBI databases

Search

Literature

- PubMed: scientific & medical abstracts/citations
- PubMed Central: full-text journal articles
- NLM Catalog: books, journals and more in the NLM Collections

Health

- PubMed Health: clinical effectiveness, disease and drug reports
- MedGen: medical genetics literature and links
- GTR: genetic testing registry
- dbGaP: genotype/phenotype interaction studies

Organisms

- Taxonomy: taxonomic classification and nomenclature catalog

Nucleotide Sequences

- Nucleotide: DNA and RNA sequences
- GSS: genome survey sequences
- EST: expressed sequence tag sequences

Genomes

- Genome: genome sequencing projects by organism
- Assembly: genomic assembly information
- Epigenomics: epigenomic studies and display tools
- UniSTS: sequence-tagged sites for genome mapping

- MeSH: ontology used for PubMed indexing
- Books: books and reports
- Site Search: NCBI web and FTP site index

- ClinVar: human variations of clinical significance
- OMIM: online mendelian inheritance in man
- OMIA: online mendelian

- SRA: high-throughput DNA
- PopSet: sequence sets
- Probe: sequence-based

- dbVar: genome structural
- BioProject: biological projects
- BioSample: descriptions of biological source materials
- Clone: genomic and cDNA clones


The screenshot shows the GQuery search interface with a grid of database categories and their descriptions. The categories are: Literature, Health, Organisms, Nucleotide Sequences, Genomes, Proteins, Chemicals, and Pathways. Each category has a list of sub-categories and their descriptions. For example, under Literature, there are PubMed, PubMed Central, and NLM Catalog. Under Health, there are PubMed Health, MedGen, GTR, and dbGaP. Under Organisms, there is Taxonomy. Under Nucleotide Sequences, there are Nucleotide, GSS, and EST. Under Genomes, there are Genome, Assembly, Epigenomics, and UniSTS. Under Proteins, there are Protein and Conserved Domains. Under Chemicals, there are PubChem Compound, PubChem Substance, and PubChem BioAssay. Under Pathways, there are BioSystems and Molecular Pathways.

<https://www.ncbi.nlm.nih.gov/nuccore/>

← → ↻ 🏠 <https://www.ncbi.nlm.nih.gov/nuccore/> ☆ ☰

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search Limits Advanced Help



Nucleotide

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

Using Nucleotide

- [Quick Start Guide](#)
- [FAQ](#)
- [Help](#)
- [GenBank FTP](#)
- [RefSeq FTP](#)

Nucleotide Tools

- [Submit to GenBank](#)
- [LinkOut](#)
- [E-Utilities](#)
- [BLAST](#)
- [Batch Entrez](#)

Other Resources

- [GenBank Home](#)
- [RefSeq Home](#)
- [Gene Home](#)
- [SRA Home](#)
- [INSDC](#)

You are here: NCBI > DNA & RNA > Nucleotide Database [Write to the Help Desk](#)

GETTING STARTED	RESOURCES	POPULAR	FEATURED	NCBI INFORMATION
NCBI Education	Chemicals & Bioassays	PubMed	Genetic Testing Registry	About NCBI
NCBI Help Manual	Data & Software	Bookshelf	PubMed Health	Research at NCBI
NCBI Handbook	DNA & RNA	PubMed Central	GenBank	NCBI News
Training & Tutorials	Domains & Structures	PubMed Health	Reference Sequences	NCBI FTP Site
	Genes & Expression	BLAST	Gene Expression Omnibus	NCBI on Facebook
	Genetics & Medicine	Nucleotide	Map Viewer	NCBI on Twitter
	Genomes & Maps	Genome	Human Genome	NCBI on YouTube
	Homology	SNP	Mouse Genome	
	Literature	Gene	Influenza Virus	
	Proteins	Protein	Primer-BLAST	

GenBank Flat File

```
LOCUS      BS3588MA1                336 bp    DNA     linear   PLN 27-JUL-2000
DEFINITION Battarrea stevenii MA-Fungi 35883 18S ribosomal RNA gene, partial
           sequence; internal transcribed spacer 1, complete sequence; and
           5.8S ribosomal RNA gene, partial sequence.

ACCESSION  AF090862
VERSION   AF090862.1  GI:6650341
KEYWORDS   .
SEGMENT    1 of 2
SOURCE     Battarrea stevenii
           ORGANISM  Battarrea stevenii
           Eukaryota; Fungi; Dikarya; Basidiomycota; Agaricomycotina;
           Agaricomycetes; Agaricomycetidae; Agaricales; Agaricaceae;
           Battarrea.

REFERENCE  1 (bases 1 to 336)
           AUTHORS  Martin,M.P. and Johannesson,H.
           TITLE    Battarrea phalloides and B. stevenii, insight into a long-standing
           taxonomic puzzle?
           JOURNAL  Mycotaxon, 67-75 (2000)
REFERENCE  2 (bases 1 to 336)
           AUTHORS  Martin,M.P. and Johannesson,H.
           TITLE    Direct Submission
           JOURNAL  Submitted (10-SEP-1998) Biologia Vegetal (Botanica), Fac. Biologia,
           Univ. Barcelona, Avda. Diagonal 645, Barcelona 08028, Spain

FEATURES   Location/Qualifiers
           source          1..336
                               /organism="Battarrea stevenii"
                               /mol_type="genomic DNA"
                               /strain="MA-Fungi 35883"
                               /db_xref="taxon:107918"
                               /tissue_type="basidiome"
           rRNA            <1..34
                               /product="18S ribosomal RNA"
           misc_RNA       35..319
                               /product="internal transcribed spacer 1"
           rRNA            320..>336
                               /product="5.8S ribosomal RNA"

ORIGIN
1  ggtttccgta  ggtgaacctg  cggaaggatc  attatogaat  agacttgatg  ggttgtcgct
61  ggctcctagg  agcatgtgca  caccogtcat  ctttatccat  ccacctgtgc  accttttgta
121 gacttggagg  catacaagca  catgcatgca  atctgggtcc  acttaoctgg  tccccaggaa
181 tggagttctg  catgggtggc  tgactctgag  gctgggtgca  gtgcgagtgc  ataccctctg
241 agtctatgtc  tttttcatac  accacatttg  catgtctcgg  aatgtattat  cacaggctgt
301 ogtgcctata  aaacacaata  caactttcag  caacgg

//
```

Header

- Title
- Taxonomy
- Citation

Features (AA seq)

DNA Sequence

EMBL Flat File

Header

- Title
- Taxonomy
- Citation

Features (AA seq)

DNA Sequence

from: Fiona Brinkman, MBB

```
ID AF115338 standard; DNA; PRO; 591 BP.
AC AF115338;
SV AF115338.1
DT 03-JUN-1999 (Rel. 59, Created)
DT 23-AUG-1999 (Rel. 60, Last updated, Version 2)
DE Pseudomonas fluorescens ECF sigma factor SigX (sigX) gene, complete cds.
KW .
OS Pseudomonas fluorescens
OC Bacteria; Proteobacteria; gamma subdivision; Pseudomonadaceae; Pseudomonas.
RN [1]
RP 1-591
RX MEDLINE; 99369842.
RA Brinkman F.S., Schoofs G., Hancock R.E., De Mot R.;
RT "Influence of a putative ECF sigma factor on expression of the major outer
RT membrane protein, OprF, in Pseudomonas aeruginosa and Pseudomonas
RT fluorescens";
RL J. Bacteriol. 181(16):4746-4754(1999).
RN [2]
RP 1-591
RA De Mot R.;
RT ;
RL Submitted (04-DEC-1998) to the EMBL/GenBank/DDBJ databases.
RL F.A. Janssens Laboratory of Genetics, Applied Plant Sciences, K.
RL Mercierlaan 92, Heverlee B-3001, Belgium
DR SPTREMBL; Q9X4L7; Q9X4L7.
FH Key Location/Qualifiers
FH
FT source 1..591
FT /db_xref="taxon:294"
FT /organism="Pseudomonas fluorescens"
FT /strain="M114"
FT CDS 1..591
FT /codon_start=1
FT /db_xref="SPTREMBL:Q9X4L7"
FT /transl_table=11
FT /gene="sigX"
FT /product="ECF sigma factor SigX"
FT /protein_id="AAD34329.1"
FT /translation="MNKAQTLSTRYDPRELSDEELVARSHTELFHVTRAYEELMRRYQR
FT TLFNVCARYLGNDRDADDVCQEVMKLVLYGLKNLEGKSKFKTWLVSITYNECITQYRKE
FT RRRRLMDALSLDPLEEASEEKALQPPEKGGGLDRWLVVVNPIDRGILVLRFVAELEFQE
FT IADIMHMGLSATKMRYKRALDKLREKFAGETET"
SQ Sequence 591 BP; 157 A; 133 C; 170 G; 131 T; 0 other;
atgaataaag cccaaacgct atccacgcgc tacgaccccc gcgagctctc tgatgaggag 60
ttggtcgcgc gctcgcatac cgagcttttt cacgtaacgc gcgctatga agaactgatg 120
cggcggttacc agcgaacatt atttaacggt tgtgcgagat atcttgggaa cgatcgcgac 180
gcagacgatg tctgtcagga agtcatgttg aaggtgctgt atggcctgaa gaacctcgag 240
gggaaatcga agttcaaac gtggctctac agcatcacgt acaacgaatg tattacgcag 300
taticggaag aacggcgaaa gcgtcgcttg atggacgcat tgagtcttga cccctcgag 360
```


LOCUS vs Accession vs PID vs protein_id: What's the difference?

LOCUS: Unique string of 10 letters and numbers in the database. Not maintained amongst databases.

ACCESSION: A unique identifier to that record (particular sequence) in GenBank/EMBL/DDBJ that does not change when record is updated.

Nucleotide gi: Geninfo identifier (gi), a unique integer specific for GenBank which will change every time the sequence changes.

VERSION: System started in 1999 for GenBank/EMBL/DDBJ where the accession and version play the same function as the accession and gi number. Format: accession.version

Which of these would you use to cite a sequence?

When would you use one over another?

- * **LOCUS:** Unique string of 10 letters and numbers in the database. Not maintained amongst databases.
- 🍅 **ACCESSION:** A unique identifier to that record (particular sequence) in GenBank/EMBL/DDBJ that does not change when record is updated.
- 🌿 **Nucleotide gi:** Geninfo identifier (gi), a unique integer specific for GenBank which will change every time the sequence changes. (*and can disappear!*)
- 🍅 **VERSION:** System started in 1999 for GenBank/EMBL/DDBJ where the accession and version play the same function as the accession and gi number. Format: accession.version



Tomato graphics from www.rottentomatoes.com

Briefly...Examples of Functional Divisions

PAT Patent

EST Expressed Sequence Tags

STS Sequence Tagged Site

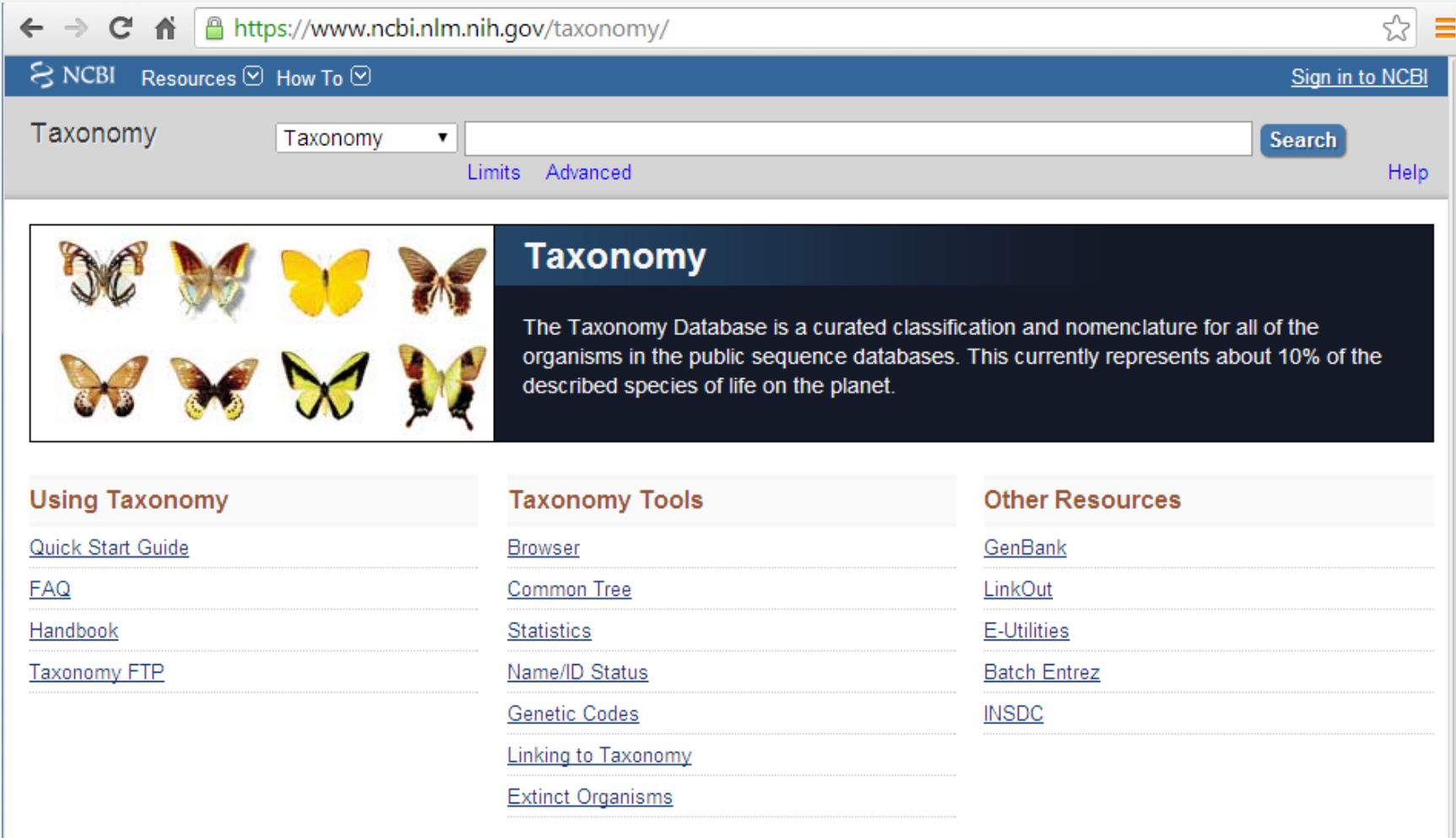
GSS Genome Survey Sequence

HTG High Throughput Genome (unfinished)

HTC High throughput cDNA (unfinished)

Genbank overview:

<http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=handbook&part=ch1>



← → ↻ 🏠 <https://www.ncbi.nlm.nih.gov/taxonomy/> ☆ ☰

NCBI Resources ▾ How To ▾ Sign in to NCBI

Taxonomy Taxonomy Search

Limits Advanced Help

Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

Using Taxonomy

- [Quick Start Guide](#)
- [FAQ](#)
- [Handbook](#)
- [Taxonomy FTP](#)

Taxonomy Tools

- [Browser](#)
- [Common Tree](#)
- [Statistics](#)
- [Name/ID Status](#)
- [Genetic Codes](#)
- [Linking to Taxonomy](#)
- [Extinct Organisms](#)

Other Resources

- [GenBank](#)
- [LinkOut](#)
- [E-Utilities](#)
- [Batch Entrez](#)
- [INSDC](#)

- Species that have not yet appeared in public entries in another Entrez database will not appear in Entrez Taxonomy (**only species in Entrez**).
- The Entrez Taxonomy database **is curated** by a small group of taxonomists at the NCBI, based on the current consensus classification in the systematic literature.
- **~10% of the total** number of described species on the planet (c. 210.000 sp.)
- It is becoming increasingly common to include some **sequence data in the description of a new species**.

Search for as lock

Display levels using filter:

Battarrea stevenii

Taxonomy ID: 107918

Inherited blast name: **basidiomycetes**

Rank: species

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](#); [Opisthokonta](#); [Fungi](#); [Dikarya](#); [Basidiomycota](#); [Agaricomycotina](#); [Agaricomycetes](#); [Agaricomycetidae](#); [Agaricales](#); [Agaricaceae](#); [Battarrea](#)

Entrez records	
Database name	Direct links
Nucleotide	24
Popset	1
Taxonomy	1

External Information Resources (NCBI LinkOut)

LinkOut	Subject	LinkOut Provider
search GBIF	taxonomy/phylogenetic	Global Biodiversity Information Facility
MA-Fungi 32476	culture/stock collections	Herbarium. Real Jardin Botanico-CSIC. Madrid
MA-Fungi 28224	culture/stock collections	
MA-Fungi 32477	culture/stock collections	
MA-Fungi 33554	culture/stock collections	
MA-Fungi 35883	culture/stock collections	
MA-Fungi 31719	culture/stock collections	
MA-Fungi 31720	culture/stock collections	
MA-Fungi 31008	culture/stock collections	
MA-Fungi 29283	culture/stock collections	
Battarrea stevenii (Libosch.) Fr. 1829	taxonomy/phylogenetic	Index Fungorum
Battarrea stevenii (Liboschitz) Fries	taxonomy/phylogenetic	MycoBank
Battarrea stevenii (taxon passport)	culture/stock collections	StrainInfo
Battarrea stevenii	taxonomy/phylogenetic	Systematic Mycology and Microbiology Laboratory. Fungal Databases

How do I update or correct errors in the Databases?

- Example: For Gene names, citations, new protein name, sequencing errors in Genbank...

update@ncbi.nlm.nih.gov

- But most people don't bother to correct things that they notice are wrong...

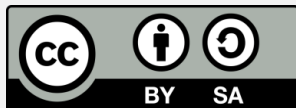
→ increased need for more focused community-based projects



from: Fiona Brinkman, MBB

Francisco Pando

Unidad de coordinación, GBIF España
Real Jardín Botánico - CSIC
Claudio Moyano 1, 28014 Madrid, Spain
pando@gbif.es
www.gbif.es



<http://creativecommons.org/licenses/by-sa/3.0/es/>

