Overview

The Entrez Taxonomy database displays the species names (and higher-level classification) of all of the organisms that are represented in the Entrez sequence databases (or any of the other Entrez databases that are indexed by taxonomy).

Species that have not yet appeared in public entries in another Entrez database will not appear in Entrez Taxonomy.

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. We try to maintain a phylogenetic classification (containing only monophyletic groups, whenever possible) but it is not automatically generated from the sequence data itself. Although there are a growing number of species with complete genomes in GenBank, the vast majority of species in Entrez Taxonomy are represented by a small snippet of sequence data, not sufficient to construct a robust phylogeny.

Searching by name

Taxonomy Entrez differs from the other Entrez databases in that the default search field is [all names] instead of [all terms] – unless you specify an explicit search field, Entrez Taxonomy assumes that you are searching with a name (or a taxid). Each entry in the Entrez Taxonomy results represents a taxon, and is identified by its 'preferred scientific name'. The 'preferred scientific name' for a taxon may be either formal (e.g. Homo sapiens) or informal (e.g. Homo sp. Altai, or uncultured fungus). Taxa above the species level may also be either formal (e.g. the order Mammalia) or informal (e.g. the unranked node eudicotyledons). In addition, each taxon may be associated with secondary names of several types – synonyms, misspellings, common names and so on. A search with the name ‘humans’ or the taxid 9606 will also retrieve the entry Homo sapiens.

Entrez currently includes sequence data from 210 thousand formally described species of eukaryotes (~10% of the total number of described species on the planet) and from 10 thousand formally described species of prokaryotes (nearly 100% of the total). The vast majority of prokaryotic diversity (and very likely the majority of eukaryotic diversity) has not yet been formally described – sequences from these undescribed species are given informal names in the taxonomy database.

It is becoming increasingly common to include some sequence data in the description of a new species. As with other papers, authors submit their sequences to GenBank to get accession numbers when they are writing their manuscript, and we see the proposed formal names prior to publication. In cases like these, we substitute informal name in the GenBank entries, and rely on the authors to let us know when their description is published in order to release the formal names. For example, taxids 906651 & 906653 were identified as Hyloxalus sp. JCS-2010a and Hyloxalus sp. JCS-2010b when they first appeared. These
were updated to *Hyloxalus yasuni* and *Hyloxalus italoi* when Paez-Vacas et al. (2010) was published.

**Boolean queries**

You can use fielded Boolean queries in Entrez to answer many questions. For example, how many amphibian taxa have appeared in Entrez for the first time this year? As with any Entrez query, it is often useful to set up “what’s new” email updates in MyNCBI for queries of interest. [Note: the ‘specified’ property flags formal names at and below the species level.]

Amphibia [subtree] AND specified [prop]

Amphibia [subtree] AND 2011 [date]


**Entrez links**

Links from Taxonomy to the other Entrez databases are typically ‘exploded’ – this means, for example, that links to Nucleotide Entrez from the taxonomy entry for the Mammalia will retrieve sequences from all of the mammalian species. There are a handful of exceptions – the default links to the literature databases (like PubMed Central) and to MeSH are unexploded. Links from the Mammalia to PubMed Central will only retrieve papers that explicitly mention the Mammalia.

**Update frequency**

There are two other ways to access the taxonomy database in addition to the Entrez Taxonomy interface – the Taxonomy Browser and the taxonomy ftp dump. Entrez Taxonomy is reindexed and updated daily, the Taxonomy Browser is updated in real time (as the taxonomy database is edited by the taxonomy group) and the taxonomy ftp dump files are updated hourly.

**Taxonomy Browser**

The Taxonomy Browser displays a hierarchical view of the taxonomy, as well as more detailed taxon-specific pages that highlight the internal links to other Entrez databases, the LinkOut links to external resources and other taxon-specific data.

http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?name=hominidae

http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?name=hominidae&lvl=0

The Taxonomy Browser also provides some useful search facilities that are not available in Taxonomy Entrez – for example, search for ‘elegans’ (without the quotes) as a ‘token set’, or search for ‘c* elegans’ (again without the quotes) as a ‘wild card’. 
Taxonomy ftp

The taxonomy ftp site includes table dumps of the contents of the taxonomy database. There is a terse readme file – the most important files are nodes.dmp (which associates each taxid with its parent taxid) and names.dmp (which associates names with taxids).