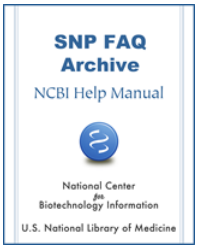
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: SNP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. Navigating in dbSNP's FTP Site. 2005 Jul 7 [Updated 2014 Feb 18].
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Navigating in dbSNP's FTP Site

Created: July 7, 2005; Updated: February 18, 2014.

FTP Site Organization

B125 changed the FTP site completely. Could you explain how the dbSNP FTP site is now organized?

The updated ftp directory is organized as follows:

The main ftp directory for dbSNP contains a number of directories, the most useful of which are:

- organisms/
- database/
- specs/

The organization of each of the above directories is explained below:

The FTP "organisms/" directory:

The FTP "organisms/" directory contains a list of the organisms for which we have SNP data, organized by common name, and followed by its NCBI taxonomy id number. For example, the organism "*Homo sapiens*" is listed as "human_9606".

Click on the organism of choice to access the ftp report files available for it. For example, if you click on human_9606 from the list of organisms, you will find that the human organism subdirectory contains the following files:

- ASN1_bin/
- ASN1_flat/
- XML/
- chr_rpts/
- rs_fasta/
- submit_format/
- genotype_by_gene/
- genotype/
- haplotypes/
- database/
- misc/

Please note that the "database/" subdirectory within a species directory (see above list) contains the following files or links to files:

organism_data
This subdirectory contains the species .bcp files for many tables, including: “Batch”, “dn_table_rowcount”, “RsMergeArch”, and “SNPAncestralAllele”.
organism_shared_data link
This link takes the user to the organism_shared_data subdirectory, which contains tables of .bcp data shared among organisms.
schema

The FTP “database/” directory:

The FTP “database/” directory contains the following subdirectories:

organism_shared_data
This subdirectory contains tables of .bcp data shared among Organisms, and includes the “Allele.bcp” file
schema
This subdirectory contains the schema of dbSNP_main, which contains tables shared among organisms.
Illumina_top_bot_strand_note
This document defines the “top” and “bottom” strands of a sequence in the absence of a reference genome to orient submissions.
b124
This subdirectory contains all the old database related files as of build 124.

The FTP “specs/” directory:

The FTP “specs/” directory contains a number of text, .pdf, .asn, and .xsd files on everything from dbSNP submission instructions, and b125 mapping information to genotype resource information and haplotype submission .xsd files. **(12/9/05)**

Is there a gradual movement toward moving NCBI data output preferentially to XML rather than ASN.1?

No, we have a large user base for both formats, and will likely support both formats in the future.(**6/8/06**)

Are there any plans to modify the XML Schema for dbSNP?

We will make future modifications to the XML schema as necessary to fit evolving SNP data model. We'll announce any schema changes on the dbSNP homepage and the dbSNP maillist, should you care to subscribe.(**6/8/06**)

I have looked through dbSNP_main tables.sql.gz, but can’t find SNPAncestralAllele.bcp.gz. Where is it?

The SNPAncestralAllele.bcp.gz for each organism is available on the FTP site in the /organism_data subdirectory of the main database directory. To get human SNPAncestralAllele.bcp.gz, you would go to the dbSNP FTP site, select “database”, then once you are in the database directory, select “organism_data”. Once you are in the organism_data subdirectory, choose “human_9606”. Once in the human subdirectory, choose “SNPAncestralAllele.bcp.gz”, located about a third of the way down the page.(**9/19/06**)

Naming Convention for SNP FTP Organism Folders

What is the difference between the dbSNP FTP folders “cow_30522” and “cow_9913”?

These folders contain the SNPs for two entirely are different organisms:

cow_30522 -- Bos indicus

cow_9913 -- Bos taurus

I know the folders on first glance don’t seem to show any apparent difference between the directories, but in actuality they do. The FTP directory name for each organism is in the format "speciesCommonName_taxonomyId", where the number portion of the name is actually the NCBI taxonomy

species level tax id. So, directory cow_30522 actually represents a Bos indicus x Bos taurus hybrid, while directory cow_9913 actually represents Bos taurus.

The problem with using specific organism names as directory names is that the scientific names for an organism can be quite long and are not necessarily obvious enough to a wide user base. Common names like "Salmon" are easy to grasp but may be not specific enough since dbSNP may contain several different species of the same organism (as is the case with salmon), so we decided to use the "speciesCommonName_taxonomyId" format as a compromise.

To help users better understand what each organism directory contains, I will propose that a "directoryName_to_organismName.README" file is placed in the dbSNP FTP site at the "/organisms" directory level to explain the directory names. This readme file would have three columns:

Directory_name | tax_id | organism_name

So for cow_30522 the entry in the README file would be:

```
Directory_name|tax_id |organism_name
Cow_30522      30522   Bos indicus x Bos Taurus
```

Until the above "README" file is in place, you can get the full taxomomic descriptions for yourself by doing the following:

1. Go to the top of the dbSNP home page where there is a text search box drop-down menu following the words "Search Entrez" and change the menu from "SNP" to "Taxonomy".
2. Put 9913 or 30522 in the text box to the left of the first text box and press "Go", and you will get a taxonomy report for it.

(**11/24/08**)

FTP Site Documentation (Docsum)

The file docsum_2005.xsd.old has an attribute called "physMapStr". The new docsum_2005.xsd does not have this attribute. Which attribute(s) in the new schema is (are) the equivalent of "physMapStr"?

The "physMapStr" attribute was deprecated (made invalid or obsolete) starting with build 125 due to a location type change. Please see the (old) Column Description for SNPContigLoc and look at the information for "phys_pos". You may also wish to look at the new location type description. The attributes that correspond to the new location type are 'leftFlankNeighborPos', 'rightFlankNeighborPos', 'leftContigNeighborPos', 'rightContigNeighborPos'.(**7/26/06**)