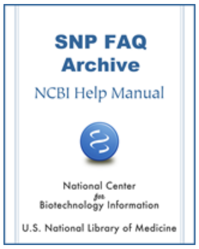
NLM Citation: SNP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. SchemaTable Definitions/Locations. 2005 Aug 8 [Updated 2010 Jun 15].
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



SchemaTable Definitions/Locations

Created: August 8, 2005; Updated: June 15, 2010.

Does each table file in the schema contain data from one database table? Am I correct in assuming that the file name is the same as the table name?

Each table file contains data from one database table only and the names match. (**12/19/07**)

AlleleFreqBySsPop

Where can I find a file that specifies the data contents and the column names of a file I downloaded from dbSNP's FTP site called AlleleFreqBySsPop.bcp.gz,?

To get a detailed description of the table, go to the dbSNP Data Dictionary (there is a link to the Data Dictionary in the Documentation section of the left blue side bar of the dbSNP home page). Once you are on the Data Dictionary site, select "Search by Table Name", place the table name in the text box and click on the "Search" button to see a general description of the AlleleFreqBySsPop table and column descriptions for that table. (**10/25/06**)

OmimVarLocusIdSNP table

What does the "var class" column in the "OmimVarLocusIdSNP" table represent? One of the numbers in the column is 10, but I don't know what it means.

The var_class is an internal code that keeps track of the variation source.

Here are the var_class label definitions:

1 (AA) : Amino acid variation (ASP156GLY)

2 (INS): 1bp base insertion

3 (NAA): Not amino acid variation (xxx67yyy)

4 (FS): Amino acid to frameshift

5 (RS): Rs link annotated by OMIM staffs

10 (UN): Unclassified variations

var_class is not supported and will probably be removed in a future release.(**11/01/07**)

Publication Table

How do I use the dbSNP "Publication" table to determine which refSNP number has been associated with a specific publication?

Look at the dbSNP Entity Relationship Diagram; although the document is rather old, page 2 and 3 are still valid. On page 2, you will see the link between Submitter, Publication, Author, BatchCita, Batch, SubSNP and SNPSubSNPLink. SNPSubSNPLink and SNP links are on page 3.

The publication table stores submitted publication information. The publication links that dbSNP mines from Pubmed will be available on dbSNP's FTPsite by build 130.

Another table you can check is the new "SubSNPPubmed" table, which stores PubMed links provided by the Clinical and LSDB submitters. You can access SubSNPPubmed.bcp.gz in the human organism_data subdirectory in the dbSNP FTP site (**05/08/08**)

SNPAlleleFreq

Why do some SNPs in the "freq" column of the "SNPAlleleFreq" table have have freq=0 or freq=1? What do these numbers mean?

In SNPAlleleFreq, when freq=0, it means that there is no occurrence of this allele. When freq=1, it means the allele occurs for all chromosomes.

Some submissions to dbSNP are genotype or frequency information for rare variations and it is possible that among the population they sampled, there was no variation. (**01/03/08**)

SNPContigLoc Table

Can I assume that rs numbers in the SNPContigLoc table, but not in the SNPContigLocusId table are intergenic regions?

There are several reasons why a SNP could be in the SNPContigLoc table, but not in the SNPContigLocusId table:

1. The SNP is intragenic.
2. The SNP has multiple hits on the genome. We only annotate those SNPs that are uniquely placed on a genome.
3. The SNP's flanking sequence is in cDNA context and has resulted in poor alignment to genome, but the SNP still could be in a gene (this case is not common).

(**06/03/09**)

I have a set of rs IDs for SNPs I want to place on build 34.3. Where do I find "b125_SNPContigLoc_b34_3"?

b125_SNPContigLoc_b34_3 is located in your organism's (human, in this case) organism_data directory in the dbSNP FTP site.

You can find the table column definitions for b125_SNPContigLoc_b34_3 in the Data Dictionary . **(1/28/05)**

Even though the SNPContigLoc table often has more than one entry per rs number, do I use it to obtain chromosome and chromosome location for a given rs or ss number?

SNPContigLoc is the right place to get chromosome and chromosome location information. When an rs maps to multiple locations, it will have several rows (entries) in SNPContigLoc. SNPMapInfo shows the number of different places (chr, contigs) an rs will map to the genome.

RefSNP page uses several procedures/views for data. Mapping data comes from the SNPContigLoc table and the SNPMapInfo table. SNP functional data (exon/intron, etc) comes from SNPContigLocusId table. mRNA mapping information comes from the MapLink table and the MapLinkPid table

Build 116 introduced some schema changes. Look at the List of changes to build115 document, the Updated dbSNP ERD, and the Data Dictionary to see documentation of these changes.

We have a program that annotates mRNA sequence with tags for SNPs or in/del (insertion/deletion) mutations. We are having a problem running the program with the in/del mutations because the in/del refSNP records in the dbSNP flat files have changed from having "^" or ".." marks in the record (our program looks for the ".." marks), to not having these marks at all. What do we do?

The change you have noticed is described on the Column Description for table: SNPContigLoc page. Look in the phys_pos column.

Where your program is looking for "..", try updating it so that it checks for ctg_start and ctg_end. ".." is used for ctg_start and ctg_end, which are not the same. Here's an example of the change:

The records used to look like this:

chr-pos=56091347..56091351 (I omitted other fields here) ctg-start=23667725 | ctg-end=23667729

Now, the above will be reported as:

chr-pos=56091347| ctg-start=23667725 | ctg-end=23667729

If your program could do the following:

If (ctg-end - ctg-start) > 0,

Then the chr-pos-range could be represented as:

(chr-pos),(chr-pos+ctg-end - ctg-start)

(**7/18/06**)

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**)

Can I determine if a SNP is uniquely mapped if I find a map location for the SNP in b126_SNPContigLoc_36_1, and if the following two mapping conditions are met: rf_ngbr - lf_ngbr - 1 = 1 and rf_ngbr - lf_ngbr = rc_ngbr - lc_ngbr ?

What you are doing is generally correct, but SNPContigLoc has mapping data for all assemblies (e.g. the human data includes both the NCBI and Celera assemblies), so you'll need to determine which assembly you have by joining with the ContigInfo table. Also, rf_ngbr - lf_ngbr - 1 = 1 may also include other types of variations, as a result of the method we use to code the variation in rs FASTA: we always use "N" for any length of indel, mixed SNP, MNP, etc. Anything that not using the the IUPAC code letters, we make an "N". As a result, many variations may fit in the two conditions you listed.

There is an online description of the SNPContigLoc table that includes an example sql showing how to determine all trueSNPs that have unique mapping on one assembly. (**8/21/06**)

How do I determine all trueSNPs that have unique mapping on one assembly?

The online description of the SNPContigLoc table includes an example sql that shows how to determine all trueSNPs that have unique mapping on one assembly.(**8/21/06**)

SNPContigLocusId Table

Not all organisms have SNPContigLocusID tables for the current build. Are the tables with the latest build numbers are the most recent versions?

Not all organisms are remapped with each dbSNP build because their genome assembly didn't change, so you should use the latest version for each organism.(**01/11/08**)

How do I determine to the refSNP (rs) ID for the snpID in theSNPContigLocusId.bcp.gz file?

The refSNP (rs) ID **is the ID** number given in the "snp_id" field of the SNPContigLocusId.bcp.gz in file. You can confirm this by doing the following:

Look at the rs number associated with a gene on a refSNP page. For example, go to the refSNP cluster report for rs268. If you look at the very top of the "Geneview" section of this report, you will see that the SNP is associated with the LPL gene (Lipoprotein Lipase). If you click on the "LPL" link at the top of "Geneview" section, you will get a Entrez Gene report for LPL, which states that the gene ID for LPL is 4023.

Now, if you go to the SNPContigLocusId.bcp file, you will see a row that contains snp_id 268. That same row will have gene_id 4023.

I should mention that you will see two rows for rs268, one representing the reference contig, and one representing the Celera contig. (**11/19/07**)

Is information on whether or not SNPs are located in the coding region or the promoter region stored in the ""fxn_class" field of "SNPContigLocusId" table?

Yes. **(3/22/05)**

Is SNP location (coding region, splice site or promoter region) stored in the fxn_class field of the SNPContigLocusId table? If so, where is the corresponding function class for promoter region?

Yes, the location information for SNPs are in the the fxn_class field of the SNPContigLocusId table, but I don't think that the current SNP function class set includes "promoter region".

Why does SNPContigLocusId.bcp.gz contain 17 fields, while the schema contains only 15 fields?

The ER Diagram is a manual effort, and as such, it may not get updated in a timely manner, so the fields may be out of sync while the update occurs.

For users who create local databases, the Schema main table is the best place to get updated schema information, as it is always updated when the FTP table data files are updated.

When we queried our local copy of dbSNP after loading the human data for build125, we found that not all of the refSNPs (rs) present in the SNP main table are represented in the B125_SNPCONTIGLOCUSID_35_1 table. Why?

Only refSNPs that are associated with a gene (via contig annotation or mRNA BLASTing) can be found in the SNPCONTIGLOCUSID table, therefore, it appears that some of the refSNPs in main table are not associated with a gene. (**4/19/06**)

SNPFunctionCode Table

I noticed that the coding non-synonymous function code has been subdivided. Can you list and define the subdivisions?

Function Code 41: "Nonsense" (coding nonsynonymous) changes to the Stop codon

Function Code 42: “Missense” (coding nonsynonymous)

alters codon to make an altered amino acid in protein product

Function Code 44: “Frameshift” (coding nonsynonymous)

indel SNP causing frameshift

You can find up-to-date function codes and their definitions in the SnpFunctionCode.bcp.gz table located in the /shared_data directory of the dbSNP FTP site. (**10/27/08**)

I noticed that b127 SNPs are no longer associated with function class codes 1,2,5, and 7. Does dbSNP no longer use these function class codes?

As of build 127, function codes 1, 5 and 7 have been modified into two digit codes that will more precisely indicate the location of a SNP. The two digit codes have function codes 1, 5 or 7 as the first digit, each of these numbers keeping its original meaning, and 3 or 5 as the second digit, indicating whether the SNP is 3’ or 5’ to the region of interest. So the new function codes are as follows:

Function code 13: “nearGene-3”

Where:
1=locus region
3= SNP is 3’ to and 0.5kb away from gene

Function code 15: “nearGene-5”

Where:
1=locus region
5= SNP is 5’ to and 2kb away from gene

Function code 53: “UTR-3”

Where:
5= UTR (untranslated region)
3= SNP located in the 3’ untranslated region

Function code 55: “UTR-5”

Where:
5= UTR (untranslated region)
5(as the second digit)= SNP located in the 5’ untranslated region

Function code 73: “splice-3”

Where:
7=splice site
3=3’ acceptor dinucleotide

Function code 75: “splice-5”

Where:
7=splice site
5=5’ donor dinucleotide

Function code 2, however, was retired permanently as of b127, as it identified a SNP as being in the coding region of a gene, but that other details about its location were unknown. Since mapping and annotation have improved dramatically since function code 2 was defined, it is no longer used. (**4/30/07**)

Why is the SNPFunctionCode table nearly empty?

SNPFunctionCode is a small lookup table of SNP functional codes, and has a grand total of 19 lines:

```
1 locus: mrna_acc and protein_acc both null. 2003-02-03 16:01:00.0 other 1
2 coding: coding 2003-02-03 16:01:00.0 cSNP  1
3 cds-synon: synonymous change 2003-02-03 16:01:00.0 cSNP  1
4 cds-nonsynon: nonsynonymous change 2003-02-03 16:01:00.0 cSNP  1
5 UTR: untranslated region 2003-02-03 16:01:00.0 other 0
6 intron: intron 2003-02-03 16:01:00.0 other 0
7 splice-site: splice-site 2003-02-03 16:01:00.0 other 0
8 cds-reference: contig reference 2003-02-03 16:01:00.0 other 1
9 synonymy unknown coding: synonymy unknown 2003-02-03 16:01:00.0 other 1
11 GeneSegment: In gene segment with null mrna and protein. ex. IGLV4-69.
   geneId=28784 2006-12-04 12:49:00.0 other 1
13 nearGene-3: within 3' 0.5kb to a gene. 2006-11-21 00:00:00.0 other 0
15 nearGene-5: within 5' 2kb to a gene 2006-11-21 00:00:00.0 other 0
41 nonsense 2005-08-01 00:00:00.0 cSNP  1
42 missense 2005-08-01 00:00:00.0 cSNP  1
44 frameshift In coding region. 2005-08-01 00:00:00.0 cSNP  1
53 UTR-3: 3 prime untranslated region 2005-08-01 00:00:00.0 other 0
55 UTR-5: 5 prime untranslated region 2005-08-01 00:00:00.0 other 0
73 splice-3: 3 prime acceptor dinucleotide 2005-08-01 00:00:00.0 other 0
75 splice-5:5 prime donor dinucleotide 2005-08-01 00:00:00.0 other 0
```

(**4/30/07**)

SNPSubSNPLink Table

The min/max build IDs for some SNPs found in SNPSubSNPLink.bcp are not consistent with those found in the dbSNP website.

“Created” on the cluster report web page is equivalent to the **min**imum or earliest build_id in SNPSubSNPLink, while “updated” refers to the SNP the most recent build_id where the refSNP cluster report page information was last updated. Since all of the refSNP report pages are updated with some new information (e.g. new mapping and/or gene annotation information or new HGVS names, or new links to OMIM, VarView etc.) every build, the “Updated Build” will be the most recent build that dbSNP has put out, and will therefore be equivalent to the **max**imum build_ID in SNPSubSNPLink. (**06/18/09**)

SubInd Table

The dbSNP FTP site has one big file: "SubInd.bcp" and single files for each chromosome called "SubInd_ch1.bcp" and "SubInd_nhm_ch1.bcp". What is the difference between these files?

The dbSNP Schema change documentation mentions that as of B127, the SubInd table was replaced with a partitioned view of a set of underlying tables by chromosomes in two separate databases to keep each database size small for ease of maintenance. You must have happened to go to the FTP site before we had a chance to clean up the old files. The updated FTP files should now be available.

Information about the SubInd table can also be found in the database dictionary by typing “SubInd” (without the quotation marks) into the text box at the top of the page, selecting the radio buttons marked “TableName” and “contains”, and clicking on the “search” button. When the response page appears, click on “SubInd link located in the TableName column to see a column description for the SubInd table. (**5/1/07**)