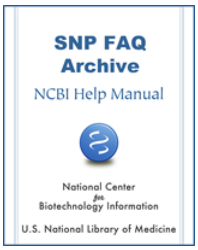
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: SNP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. Locating Schema Tables that Contain Specific SNP Data. 2005 Aug 8 [Updated 2010 Jun 15].
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Locating Schema Tables that Contain Specific SNP Data

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

Double Hit Data Table

Which tables contain SNP double-hit information?

The SNPVal table contains the rs numbers from validation submissions, and has two columns: batch_id and snp_id. To get the entire double hit list that Dr.Jim Mullikin provides to dbSNP, use all the rows in SNPVal that have batch_id = 11597. If you use all the rows in SNPVal that have batch_id =7838, you will get some double hits that dbSNP mined.

Functional (synonymous, non-synonymous, etc.)Data Table

If a SNP is located in an exon region, what column tells me whether or not the SNP is a synonymous or a non-synonymous change?

The column FXN_CLASS reports the functional class for the SNP. You can find the column definitions in the Data Dictionary. (search for SNPContigLocusId).

If you want functional data for a just a subset of data, I would suggest trying our batch query service and select FLATFILE format. The FLATFILE format is a human readable format which is defined in an FTP document called "FTP resources for the dbSNP Database". The LOC lines report the functional data for each SNP.

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 there information in the table headers that indicate whether or not a SNP is located in an exon (coding), an intron or a UTR region?

The SNP functional data are organized by organism and stored in tables with the base name SNPContigLocusId (human data). Prefix 'mm' and 'rn' for mouse or rat data, respectively.

The table is keyed on rs number and reports the SNP x mRNA feature functional class data based on either SNP location (locus region, UTR, intron, splice site) or location+SNP allele (coding synonymous, nonsynonymous, contig reference).

What is the name of the table that contains the dbSNP's "OMIM" and "function" data?

The data you want is located in tables SNP2Omim and SNPContigLocusId. Please see the Data Dictionary for table column descriptions.

Genotype Data Tables

For a given SNP, where do I find the total number of individuals with genotype data in the dbSNP tables?

Individual counts with genotype data for a refSNP is not directly stored in a table. The web page uses a view that looks at the SNPSubSNPLink table, the SubInd table, and the SubmittedIndividual table.

See the following example view:

```
create view vwSNP_ind_cnt
as
select l.snp_id, count(distinct i.ind_id) ind_cnt
from dbo.SNPSubSNPLink l, dbo.SubInd s, dbo.SubmittedIndividual i
where l.subsnp_id = s.subsnp_id and s.submitted_ind_id =
i.submitted_ind_id
group by l.snp_id
go
```

A data dictionary that includes the definitions for the above three tables and columns is in progress as of this writing (**7/03**).

Is there a document that describes the content of the mouse genotype tables in the dbSNP FTP site?

The dbSNP schema documentation, should answer your question. **(4/8/05)**

Mapping Data Tables

Where in the FTP site do I find a table with coordinates of refSNPs (or subSNPs) in the genome (build 36)?

You can find the latest chromosome position of a refSNP in the phys_pos_from column of the b129_SNPContigLoc_36_3 table. (**07/10/08**)

Which table provides the sequence position of a SNP on a gene?

As far as I know NCBI doesn't have a canonical set of "gene records" that would provide a coordinate system for SNP alignment to genes. NG_ records are curated gene regions, but they do not exist for all genes annotated on the human genome. In any event, we make no systematic effort to map our SNPs to this set apart from NCBI genome alignment.

dbSNP maintains a table called ContigExon which provides the start and stop positions of each exon in contig coordinates for each mRNA refseq defined in a particular NCBI genome build. We use these coordinates to create the exon placement graphic on our Geneview display.

You can obtain the same canonical list of exon boundary information in chromosome coordinates from an NCBI genomes ascii file. **(4/20/05)**

Where can I download files that describe the mapping relationships between human submitted SNPs (ss) and human RefSNPs (rs)?

You can get ss to rs mapping using the SNPSubSNPLink.bcp.gz table file, located in your organism's (human, in this case) organism_data directory in the dbSNP ftp site. You can get the field names of the SNPSubSNPLink table from the dbSNP_main_table.sql.gz file, located in the shared_schema directory of the SNP ftp site. **(4/29/05)**

How do I find a coordinate system based on the underlying clones used in an assembly?

If what you want is the chromosome coordinate system, we have a column in SNPContigLoc for that. The integers located in SNPContigLoc.phys_pos_from record the start position of each SNP, but in a few cases

this is left NULL because the position and/or orientation of the underlying contig on the chromosome is not known. We also record an ascii string which is used for display purposes in SNPContigLoc.phys_pos. phys_pos to make it easier to work with.

When a range or a deletion is involved, it can take one of the three following forms:

12345	a true SNP mapping to a single base
12345^12347	mapping as a deletion on the chromosome. usually an INDEL
12345..12367	mapping as a range on the chromosome usually an INDEL or microsatellite.

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**)

Merged/Expired RefSNP (rs) Tables

How does the RsMergeArch table relate to the SNPHistory table?

Let me explain how we track both merged and deleted (an entirely different process from merging) refSNP (rs) numbers, by using a hypothetical example where a "chain merge" (multiple rs numbers merge into each other) occurs:

For example, let us say rs "A" merged into rs "B", and later, rs "B" merged into rs "C". As a result of the first merge, the entry for rs "A" in rsCurrent

is updated to rs "B"; after the second merge, rsCurrent is then updated to rs "C". Now, if rs "C"'s submitters withdraw all the member ss numbers within the refSNP cluster rs "C", then rs "C" will get an entry in the SNPHistory table (the SNPHistory table ONLY contains SNPs that have "become history" — that is, SNPs that have been completely deleted). Please see ftp file for SNPHistory.bcp (located in the snp/database/organism_data/species of interest directory). To find the column names for the SNPHistory table, download the human_9606_table.sql, which is located in the human organism_schema directory.

Getting back to RsMergeArch: since "withdrawing rs "C" is not a merge action, the table RsMergeArch is not updated. RsMergeArch is used to track "rs merge" actions only. I can see that this might be confusing, so when time allows, we will add the following explanation to the RsMergeAch table definition, to make the RsMergeArch.rsCurrent meaning clearer:

RsMergeArch is used to track each rs merge event.

If an rs number in RsMergeArch.rsCurrent is withdrawn from dbSNP by submitter request, then an the rs number of the same value as that in rsCurrent will be entered into the SNPHistory table (which contains deleted rs numbers only).

Please note: "rsCurrent" in RsMergeArch does not mean the "current rs number" in the current dbSNP build". (**08/12/08**)

rs4823903, which has merged into rs4253690 (see RsMergeArch) still appears in SNPChrPosOnRef

Cases like rs4823903 have a bit of a twist in their history that you can see in SNPHistory:

```
 [502] MSSNP_LOAD.human_9606.2> select * from SNPHistory where snp_id =4823903
[502] MSSNP_LOAD.human_9606.3> go -mvertsnp_id:              4823903

create_time:          Feb 20 2003 01:26PM
last_updated_time:    Sep 07 2006 01:03PM
history_create_time: Oct 25 2006 12:19PM
comment: Re-activated:PHARMGKB:rs4823903|b126->rs60186231|b129:NT_011523.11_1851628
```

Submitter PHARMGKB submitted many high visibility SNPs in 2003, but also withdrew a large batch in 2006. Many of their SNPs are still valid and have been cited and used successfully in experiments by dbSNP users. PHARMGKB has since re-submitted many of their withdrawn SNPs to dbSNP, and we decided to "re-activate" the same rs numbers for those SNPs that have the same flanking sequences.

The "re-activation" was noted in the SNPHistory.comment field. For those re-activated SNPs that merged, I did not go back to change the merge history since it was an event that had indeed occurred. I can see that more work need to be done to clarify the multiple sequence of events for these SNPs. For users who actually use data from RsMergeArch and SNPHistory such clarification is necessary. (**09/09/08**)

Where do I find a list of merged RS IDs?

The merge history is in the RsMergeArch.bcp table:

Example:

```
zgrep 12753158 RsMergeArch.bcp.gz
12753158        7539545 123     0       2004-09-24 18:47:00
 2004-10-10 12:15:00     7539545 0
```

The RsMergeArch table description can be found by using the table name search option in the dbSNP database dictionary.

You can also retrieve a list of merged rs numbers from Entrez SNP. Just type “mergedrs” (without the quotation marks) in the text box at the top of the page and click the “go” button. Each entry in the returned list will include the old rs numbers that has merged, and the new rs number it has merged into (with a link to the refSNPpage for the new rs number). You can limit the output to merged rs numbers within a certain species by clicking on the “Limits” tab and then selecting the organism you wish from the organism selection box. (**2/27/06:11/03/08**)

Where do I find the tables storing expired rs id lists (merged, withdrawn, etc.)?

The tables you require are:
RsMergeArch.
SubSNPDeletedBySubmitter
SNPSubSNPLinkHistory
SNPHistory
SubSNPDelComm

For table descriptions, see the Data Dictionary

Population Diversity Tables

Which tables and columns in the schema contain the data found in the Population Diversity section of the refSNP Cluster Report?

The three values below, found in the refSNP Population Diversity section, are from tables in the SNP annotation area of the schema pdf:

Hardy-Weinberg Probability is located in the SNPHWProb table.
Average estimated heterozygosity is located in the SNP table.
Average Allele Frequency is located in the SNPAlleleFreq table.

Assay sample size (number of chromosomes), Population data sample size (number of chromosomes), Total number of populations with frequency data, and Total number of individuals with genotype data are all located in the refSNP page Population Diversity section, and are all computed on the fly.

Split RefSNP (rs) Table

I need the refSNP (rs) split list and thought that it would be in RsSplitArch and RsSplitArch_ssLink, but found that both files were empty. Where is the rs split list?

The RsSplitArch table is the reverse of the RsMergeArch table; it currently has 70 rows, and its content is based on submitter feedback regarding our rs cluster merging results. For example, rs13361 was merged into rs6154 on 4/21/2003. Later, the submitter of both rs numbers reported that the two variations were truly different and should not be merged. As a result of this comment, we split rs6154 back to rs13361 and rs6154, removed the rs1336 to rs6154 rows in RsMergeArch table, and then saved this "split or merge reversal" fact in the RsSplitArch table.

Rs split processing is still not integrated or automated at this time. The split data is located in an internal dbSNP database. If you are interested in obtaining the list of the 70 rs cluster splits, let dbSNP know, and we can email the file to you. (**2/27/06**)

STS Accession Data Table

What table is used to store dbSTS accessions?

dbSNP.SubSNPAcc.acc_part stores dbSTS accessions when SubSNPAcc.acc_type_ind = 'S'.

For example:

```
select top 10 * from SubSNPAcc
where acc_type_ind = 'S' and acc_part like '[ZG]%'
--( Most dbSTS accession starts with "G" or "Z".)

subsnp_id      acc_type_ind    acc_part         acc_ver
---------      ------------    --------         -------
4206                S           G01792           NULL
4020                S           G01913           NULL
4021                S           G01976           NULL
2304                S           G02363           NULL
2381                S           G02365           NULL
743                 S           G02368           NULL
2297                S           G02380           NULL
2298                S           G02380           NULL
19462               S           G02380           NULL
```

```
2170                S           G02394          NULL

(10 row(s) affected)
```

SubSNPAcc.acc_ver is null in this case, because most dbSTS accessions do not have a version number

Submission Data Tables

To find the submitted population information for a submitted SNP, I did the following: select subsnp_id, batch_id from subsnp where subsnp_id =2984380, but the batch_id's do not line up, and the data_dictionary says I have two foreign keys. What am I doing wrong?

SubSNP and SubPop data are submitted in different batches, with the batch_type beginning 'SNP' and 'POP' respectively.

TSC-CSHL submits over 10K in population data under new batch IDs (batch type POP) for existing ss numbers that were submitted earlier under various batch_id of batch type 'SNP'. Take a look at this example on dbSNP.

To get population information for which the submitted SNP has frequency, given a submitted SNP, joining on the ss number is enough. You don't need to go through Batch. Also, note that the population text is in the PopLine table. Try using the following SQL on ss number

14759:

```
select distinct p.pop_id, pl.line_num, pl.line,p.handle
from SubPop sp
join Population p on sp.pop_id = p.pop_id join PopLine pl on p.pop_id = pl.pop_id
where sp.subsnp_id = 2984380
order by p.pop_id, pl.line_num
```

This SQL returns:
pop_id: 738
line_num: 0
line: 752 anonymous unrelated Japanese volunteers
handle: YUSUKE

pop_id: 738
line_num: 1
line: Nation:Japan
handle: YUSUKE

(2 rows affected)

The data dictionary says that SubSNP.batch_id references Batch.batch_id; SubPop.batch_id references batch_id; and SubSNP.subsnp_id references SubSNP.subsnp_id. The ER diagram makes this relationship much more clear.