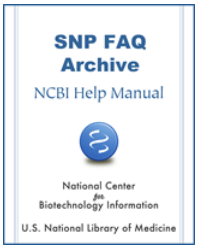
U.S. National Library of Medicine
National Center for Biotechnology Information

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



Local Copies of dbSNP

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

I am building a local copy of dbSNP, but am finding that there are a lot of files without a corresponding table in your schema — what do I do?

The database is in constant development and often there are more table files than described in the schema. You only need those tables that are mentioned in the latest schema. (**12/19/07**)

Direct Dumps of dbSNP

Can I get a direct dump of the entirety of dbSNP? Building a local copy using the FTP tables is too hard and time consuming.

Due to security concerns and vendor endorsement issues, we cannot provide users with direct dumps of dbSNP. The task of creating a local copy of dbSNP can be complicated, and should, therefore, be left to an experienced programmer. (**4/24/06**)

Creating a Local Copy of dbSNP for the Beginner

Although I have very little database experience, I must create a local copy of dbSNP. Is there a good resource for advice on how to do this?

Although it may be possible for a beginner to make a local copy of dbSNP using the schema and data from our ftp site, an experienced programmer who knows SQL and can write complex queries will be required to actually use the local database. In my experience at dbSNP, I have found that users who make local copies of dbSNP usually have dedicated programmers.

With that said, if you still intend to create a local copy of dbSNP, take a look in the "How to create a Local copy of dbSNP" section of the NCBI handbook.

You might also want to check out NCBI's eutil/efetch utilities that allow you to automate data retrieval from the web, as well as NCBI workshops for users wanting to learn data retrieval techniques. **(11/21/05)**

Is it possible for us to create a mirror of your database on our server?

Please see the instructions for creating a local sql dbSNP database located on dbSNP FTP site. **(5/17/05)**

Creating a Local Copy of dbSNP in Various Database Systems

We are creating a local copy of dbSNP on MySQL, but find that MySQL treats data as case insensitive, and your data seems to be case sensitive. Could you alter your data so that it will work with MySQL?

dbSNP uses case sensitive data and MySQL treats data as case insensitive. Generally speaking, dbSNP tries to provide a platform neutral format that will work with most databases, and as such we do require the user to find solutions that will allow the data to work with a particular server.

In this case, however, I do have the information that should allow you to set up My SQL with “case sensitive collation:

```
== Case insensitive collation

mysql>  select @@character_set_server, @@collation_server;
+------------------------+--------------------+
| @@character_set_server | @@collation_server |
+------------------------+--------------------+
| latin1                 | latin1_swedish_ci  |
+------------------------+--------------------+
1 row in set (0.00 sec)

mysql> create table foo(bar varchar(10));
Query OK, 0 rows affected (0.00 sec)

mysql> insert foo(bar) values ('aaa'), ('AAA');
Query OK, 2 rows affected (0.00 sec)
Records: 2  Duplicates: 0  Warnings: 0

mysql> select bar from foo where bar='aaa';
+------+
| bar  |
+------+
| aaa  |
| AAA  |
+------+
2 rows in set (0.00 sec)


== Case sensitive collation

mysql>  select @@character_set_server, @@collation_server;
+------------------------+--------------------+
| @@character_set_server | @@collation_server |
+------------------------+--------------------+
| latin1                 | latin1_general_cs  |
+------------------------+--------------------+
1 row in set (0.00 sec)

mysql> create table foo(bar varchar(10));
Query OK, 0 rows affected (0.00 sec)

mysql> insert foo(bar) values ('aaa'), ('AAA');
Query OK, 2 rows affected (0.00 sec)
Records: 2  Duplicates: 0  Warnings: 0

mysql> select bar from foo where bar='aaa';
+------+
| bar  |
+------+
| aaa  |
+------+
1 row in set (0.00 sec)
```

```
mysql> select bar from foo where bar='AAA';
+------+
| bar  |
+------+
| AAA  |
+------+
1 row in set (0.00 sec)
```

(**06/25/09**)

I am creating a local copy of dbSNP, and was wondering if you are using Sybase as your database.

No. Our current db is not Sybase. Our database server is:

Microsoft SQL Server 2000 - 8.00.2039 (Intel X86)

May 3 2005 23:18:38

Copyright (c) 1988-2003 Microsoft Corporation

Standard Edition on Windows NT 5.2 (Build 3790: Service Pack 1)

(**11/16/06**)

We are creating local copy of dbSNP and want to know where the corresponding DDL files for Oracle are on the dbSNP site.

We do not have Oracle specific DDL, but most of our table creation DDL statements are compliant with ANSI SQL92 standard. I suggest you try ERWin. It should be able to convert DDL to the Oracle format. (**3/1/06**)

Creating a Local Copy on Linux

I want to install a local copy of dbSNP on Linux, but as MSSQL can't be installed on a Linux platform, I want to use PostgreSQL — how do I format the *.sql file in PostgreSQL format?

We used to have similar problems with loading MySQL dumps. Since converting DDL/DML statements proved to be problematic, we end up installing a MySQL server, loading the dumps in there, and thenpulling the data into MS SQL using DTS with MySQL ODBC driver.(**10/16/08**)

Creating a Local Copy that Contains Specific SNP Data

I've built my own local SNP database, and want to add the "average allele frequency" data to my dataset — where can I get it?

You will need the SNPAlleleFre.qbcp.gz and the Allele.bcp.gz tables located in the dbSNP ftp site. The above links go to the human tables, but if you substitute the name of your organism of interest in for "human_9606" in the URL, it will take you to the table you need. **(12/12/05)**

Creating a Local Copy that has Specific Applications

Can you tell me if it is possible to create a local copy of dbSNP that has visualization and data upload interfaces as well as the ability to submit to dbSNP?

Yes, technically it is doable, but not a small feat. You can download dbSNP data from the ftp site to create a local database and program an interface to it. The "Local Copies" of this archive's "Schema" section has a lot of useful information as does the dbSNP Handbook.(**05/19/08**)

Keeping Local Copies Updated

Do I need to rebuild my database with a new backup every time there is a new build?

If you want to synchronize your database with dbSNP's latest build, you will need to rebuild your database.

Are new SNPs added to dbSNP in between dbSNP builds? We are mirroring dbSNP on our campus and need to know how often we should be updating to stay current.

In general, new SNPs are added to dbSNP during each build. So it will be necessary to update your local database every time dbSNP releases a new build if you want it to stay current.

That said, there might be some rare instances when a dbSNP new build has no new SNPs, just new frequency data, genotype data, or new mapping information. Please check http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi when there is new build announcement for more details.

Problems Downloading Data for Local Copies

I just created a local copy of human dbSNP, but my routine failed to load files such as SubSNPSeq3_p2, SubSNPSeq3_p2_human and the like. Are the files broken down into subsets?

You are on the right track. Human SubSNPSeq5 has 4 parts (see below) below. SubSNPSeq3. SubSNPAcc and SubSNPCommLine are parallel.

```
/*----------------------------
sp_helptext SubSNPSeq5
----------------------------*/
Text

create view SubSNPSeq5
as
select * from dbSNP_sup..SubSNPSeq5_p1_human
union all
select * from dbSNP_sup..SubSNPSeq5_p2_human
union all
select * from dbSNP_sup..SubSNPSeq5_p3_human
union all
select * from SubSNPSeq5_ins
```

I will see if we can put out the view text for horizontal partitioned views on the ftp site so as to clarify this in the future. (**11/08/07**)

I'm trying to create a local copy of dbSNP, and installed dbSNP_main_table.sql, and then installed dbSNP_main_view.sql. The main table installs well, but I get errors when I install the view file.

Skip all the views in dbSNP_main. They are there for dbSNP internal convenience and should not be exported. For example, vw_all_Batch is a view that is a union of all Batch tables from all organisms in dbSNP. When installed, this view requires that all the other organism databases contained in dbSNP must be created in your local copy too. Since most users are only interested in a few organisms, this view is not necessary. We will work on clean up such "non-essential" views in the FTP files to avoid future confusion. **(10/27/06)**

I've tried downloading FTP files to create a local copy of dbSNP, but get an error message that reads: "Passive mode refused. Turning off passive mode. No control connection for command. No such file or directory". What do I do?

This message indicates that your firewall is blocking passive FTP — at least some of the time. Try a "smart" FTP client like NCFTP (often available on most UNIX machines). It is better at auto-negotiating active/ passive FTP connections than some of the older FTP clients (e.g. SUN SOLARIS FTP). (**4/25/06**)

I followed your procedure for creating a local copy of dbSNP, but could not find the <table>.bcp files in the /shared_data/ directory.

All the files with the extension <table>.bcp.gz are compressed, so you'll have to uncompress them. There are online instructions that will guide you through this process. (**8/31/06**)

Using SQL to Extract Data

What is the SQL for finding the minor allele frequency (population frequency) given a SNP, and what is the SQL for finding the chromosome mapping location for a SNP?

Your questions indicate that you are familiar with the minutiae of the dbSNP schema and have a local copy of dbSNP, so I will provide SQL examples that should answer your questions:

SQL Example 1: Finding minorAlleleFreq for a set of rs# (< 100) by population:

```
select mf.snp_id as rs, f.subsnp_id, p.handle, p.loc_pop_id, cast
(mf.minFreq as numeric(3,2)) as minorFreq, a.fwd_allele from
AlleleFreqBySsPop f join  ( select l.snp_id, l.subsnp_id, p.pop_id,
min(freq) as minFreq, sum(cnt) as samplesize ,
l.substrand_reversed_flag  from  SNPSubSNPLink l   join AlleleFreqBySsPop
 p on p.subsnp_id = l.subsnp_id and l.snp_id < 100  group by l.snp_id,
 l.subsnp_id, l.substrand_reversed_flag, p.pop_id ) mf on f.subsnp_id =
 mf.subsnp_id and f.pop_id = mf.pop_id and f.freq = mf.minFreq and
f.freq != 1 join dn_Allele_rev a on a.allele_id = f.allele_id and
a.rev_flag = mf.substrand_reversed_flag join Population p on p.pop_id
 = mf.pop_id order by mf.snp_id, f.subsnp_id
```

SQL Example 2: Finding the chromosome position of weight 1 SNPs on NCBI 'reference' assembly for rs221 and rs332. Note that rs221 is trueSNP, rs332 is an indel:

```
Select m.snp_id, m.phys_pos_from+1 as snpChromPosStart,
rc_ngbr+i.contig_start as snpChromPosEnd From b125_SNPContigLoc_35_1
 m join b125_ContigInfo_35_1 i on i.ctg_id = m.ctg_id and
i.contig_label='reference' Join b125_SNPMapInfo_35_1 w on w.snp_id =
m.snp_id and w.assembly = i.contig_label and w.weight = 1  Join SNP s
on s.snp_id = m.snp_id Join UniVariation u on u.univar_id = s.univar_id
where s.snp_id in ( 221, 332)
```

(**9/5/06**)

I'm attempting to create a local copy of dbSNP. The Unix C shell script you provide on the dbSNP FTP site refers to the command: "dsql". What is "dsql" and where can I obtain it?

Thank you for your question. "dsql" is a SQL query tool for MySQL, Oracle, Postgres, MS- SQL, etc. You should be able to find it using google.

If you are using a local database, you probably have some type of SQL query tool. For example, I am currently using sqsh-ms as a command line query.

The NCBI handbook chapter for dbSNP contains detailed information for those who wish to create a local copy of dbSNP. There are also a number of dbSNP FAQ Archive sections that deal with the creation, maintenance, as well as general questions about local copies of dbSNP. (**12/03/07**)

The NSE Tables

Is there a way of viewing NSE.dtd with the details of the mod files integrated?

The NCBI_Entity.mod contains the definitions of the most common types used in NSE.dtd. The schema itself is in NSE.mod. The NSE.dtd contains essentially the directives, so if you just merge the two *.mod files, I think you'll get the integration you want. **(1/25/05)**

Searching Local Copies of dbSNP

Is there a MySQL friendly version of dbSNP somewhere for creating a local copy? Is there an optimal way of querying a local copy in MySQL format —mine is slow.

The FTP downloads are generic tables that should work in any database, so there isn't a specific MySQL version of dbSNP. We don't have experience with MySQL so are unable to offer you much help. Here are some suggestions:

Hardware

- Use a dedicated and fast computer
- Add more memory (you should have at least 4 GB)

Fast storage and disk access

- Use a query analyzer to profile where the bottleneck is in your sql, although I'm not sure if there a query analyzer for MySQL.

(**07/15/08**)

I'm working with a local copy of dbSNP and have extracted all the SNPs which affect amino-acid changes. How do I determine which amino acid is wildtype?

There is no designation of a "wild-type" allele since the occurrence of the common allele is population specific. You will have to check the frequency data over each specific population to infer whether the allele is wild type or a mutant in a specific population. Take a look at the description for the SubPopAllele table. (**03/25/08**)

Tools for Searching Local Copies

Do you have tools or NCBI utilities available that will allow me to interface and search a local copy of dbSNP?

We don't have utilities for searching local copies of dbSNP; we only have utilities that allow users to interface with NCBI's dbSNP. (**3/3/06**)

Discrepancies between a Local Copy Search and a Web-based Search

If I query Entrez for all SNPs within the first 6000 positions of chromosome 1, I get 11 hits. However, if I do the same search our local copy of dbSNP, I get 115 hits. Why is there a difference? How do I query Entrez and get the results I got from the SQL search, and vice versa?

The difference in result yield between an Entrez search and a direct search on you local database is due to the fact that Entrez only returns SNPs with clear mapping to the current genome build.

To get the same results that you got from Entrez by using a direct SQL query, use an example sql (located in the data dictionary entry for SNPContigLoc), which allows you to find all refSNPs that have unique mapping positions.

Getting the same results that you got from a direct SQL query by using Entrez is not possible right now. Before I explain, please note the definitions for mapping weight available online.

If you use your use your SNPContigLoc example (mentioned above) to make a direct query of your local database for a SNP with hits on chromosome 1 in the 1:6000 bp range, you will find all SNPs (including high mapping weight SNPs). The problem is, though, if you do an Entrez search for the same, you will only find mapping weight 1 SNPs for chromosome 1 in the 1:6000 range, since Entrez separately indexes mapping weight 2 SNPs in the chrMulti table, and does not index location. So searching Entrez using 1:6000 will not find those SNPs that have a mapping weight of 2. (**08/09/07**)

Discrepancies between the Data Dictionary and FTP Documentation

Your descriptions of mapweight in the FTP README for Chromosome Reports and for SNPMapInfo in the data dictionary are contradictory. Which is correct?

The definitions of map weight for chromosome reports and database tables are indeed different:

Chromosome Reports	Database Tables
Mapweight 1 = Unmapped	Mapweight 1 = SNP aligns exactly at one locus
Mapweight 2 = Mapped to single position in genome	Mapweight 2 = SNP aligns at two locus on same chromosome
Mapweight 3 = Mapped to 2 positions on a single chromosome	Mapweight 3 = SNP aligns at less than 10 locus
Mapweight 4 = Mapped to 3-10 positions in genome (possible paralog hits)	
Mapweight 5 = Mapped to >10 positions in genome	Mapweight 10= SNP aligns at more than 10 locations

The mapweight definitions for the database are different for historical reasons, so both definition series are correct: The mapweights defined in the chromosome reports section of the FTP README are true for chromosome reports, and the definitions given for the database tables in the database dictionary are true for all FTP data table files. (**07/14/08**)