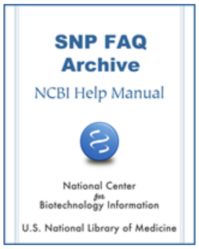
NLM Citation: SNP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. SNP Report Files. 2005 Jul 7 [Updated 2014 Feb 18].
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



SNP Report Files

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

dbSNP FTP Site Report Formats

ASN.1. and ASN.1 Flat Files

Data Discrepancies between dbSNP's Web Site and ASN.1 Files

The dbSNP website has heterozygosity information for rs1800206, but the chr22 ASN1_FLAT files show heterozygosity as null (het=?). How do I get complete information for every SNP without having to parse XML?

Consult the Data Dictionary to determine which table column holds the heterozygosity data or check the dbSNP Main Table.

Be aware that snp_type is no longer in the table because we never used it.

Functional class is located in the SNPContigLocusId table. If you need help deciphering the table, please consult the Data Dictionary mentioned above.

If you are a power user that regularly downloads dbSNP for in-house analysis, you might be interested in the ER (Entity Relationship) diagram, which delineates dbSNP structure, including the relationship between the database tables.

The refSNP page for rs28928880 shows the refSNP's amino acid position to be 25, but b126_SNPContigLocusId_36_1.bcp file for human shows the amino acid position of rs28928880 as 24.Why are these data different?

The sequence coordinate data for the XML, ASN.1, .bcp, and the Genotype/genotype_by_gene files were changed from 1-based to 0-based starting with dbSNP build 125. The ASN.1_flat, Chromosome Report, and the web page reports remain 1- based. (**6/30/06**)

ASN.1. Files Missing Data

I downloaded human XML and ASN reports for build 125, but found that many of the SNPs in these reports do not have population frequency data.

Some submitters did not submit genotype or frequency data to dbSNP in their submissions; therefore, there is no population frequency data for these SNPs. There are approximately 27 million submitted SNPs in dbSNP, and only 3.5 million of those have frequency data associated with them. (**1/9/06**)

ASN.1. Decimal Format

Since the estimated heterozygosity field in the ASN.1 format does not have a decimal point before or after zeroes (e.g. het [410112023353577, 10, -15]), how do I distinguish between 0.2, 0.02 and 0.002?

The ASN.1 format for floating point numbers represents the mantissa (the fractional part of a logarithm, to the right of the decimal point) as a whole number that includes the base number as well as its order of magnitude. In your case, the number would be base (410112023353577) and exponent 10^(-15). **(6/8/05)**

ASN.1. flat Report Definitions

Can you provide an in depth description of the SEQ line in the ASN.1 Flat file?

You can find the description in the FTP 00readme file:

```
KEYWORD       docsum.asn field (Value(s) or definition(s))
SEQ               1. PrimarySequence.attlist.gi
                  2. PrimarySequence.attlist.source
                 (submitter = sequence source is dbSNP submitter)
                 (blastmb = sequence source is NCBI blast database)
                 (xm = sequence source from NCBI Refseq)
                  3. MapLoc.attlist.asnFrom   [../^][MapLoc.attlist.asnTo]
                 (see** below for symbol definitions)
                  4. MapLoc.attlist.locType   (1=insertion; 2=exact;
                  3=deletion; 4=range-insertion; 5=range-exact;
                  6=range-deletion)
                  5. MapLoc.attlist.orient (+ =forward, - =reverse, ?
                  =unknown)
```

(**08/07/08**)

Movement to XML rather than ASN.1 Output?

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**)

Chromosome Report Files

Why is allele data not included in chromosome reports?

The chromosome report format is a condensed report of SNP mapping and validation properties. Allele data is strand specific and must be taken in context. You can get allele information for a SNP by using the ASN.1 flat file format , located on the dbSNP ftp site, for your organism of interest (human in this case). **(3/24/05)**

Where can I find a description for chr_rpts? I want to know which column in chr_rpts represents the contig orientation of a SNP hit.

The description of the chr_rpts files are in the dbSNP FTP readme file.

The Chr_rpt column definitions are located about three-quarters of the way down the FTP readme file under the “CHROMOSOME REPORTS” section heading. Although SNP orientation is not reported in the chr_rpts files, you can find SNP orientation by looking at the entry for a specific refSNP(rs) number in the ASN1_flat files. Look for SNP orientation in the CTG line of the entry for the rs number of interest. Below is an entry for an rs number taken from the ASN_flat files:

```
rs8896|human|9606|snp|genotype=NO|submitterlink=YES|
|updated 2004-10-04 13:37|ss10932|CGAP-GAI|52782|orient=+|
|ss_pick=YES SNP|alleles='C/T'|het=?|se(het)=? VAL|validated=NO|
|min_prob=?|max_prob=?|notwithdrawn CTG|assembly=reference|
|chr=MT|chr-pos=8270|NC_001807.4|
|ctg-start=8270|ctg-end=8270|loctype=2|orient=
```

(2/14/06)

I need to download all the SNP chromosome positions in build 125 and 126, but the chr_rpts files I downloaded were the same for each build I selected.

I assume you want human build 125 and 126 map positions, right? You could get them in the organism_data subdirectory for each organism (the link above goes to the human organism_data subdirectory). **Please Note**: the files you are looking for start with the build number followed by "SNPContigLoc"; e.g. b125_SNPContigLoc_35_1. The column definitions for SNPContigLoc are located in dbSNP's Database Dictionary. (**03/07/08**)

FASTA Files

How do I get an incremental update of my FASTA files?

dbSNP does not provide incremental updates. You'll have to download a complete new set of FASTA files from the FTP site.

I downloaded all 4 fasta files from the Macaque FASTA file on the dbSNP FTP site, but only the Q2_ss.fas.gz file has data in it. Why don't the other 3 files have any data in them?

Our automatic script generates the submitted SNP (ss) fasta reports using the submission date for each quarter as the file title, so, the fasta files are labeled for each quarter: Q1, Q2, Q3, and Q4. An empty file means that we didn't have any submission in that particular quarter.

(**7/18/06**)

I want to download the current set of sequences in dbSNP to format for a local BLAST search. Could I use a "wget --mirror" to pull down the contents of /snp/organisms/*/ss_fasta/*/*", and then concatenate the files?

We already have the files you need on the dbSNP FTP site as FASTA files. For example, to get human ss FASTA data, go to the human organism directory, and select ss_fasta,and then select the year in which the data you need was submitted.

You can also get rs FASTA organized by chromosome in the dbSNP FTP site, or you could also blast dbSNP rs sequences directly online.

For more information, please check the dbSNP handbook. (**9/13/06**)

Can you send me the FASTA sequence for a list of rs numbers?

You can use dbSNP Batch Query Service to download the FASTA format for a list of rs numbers. (**9/19/06**)

XML Files

Data Discrepancies between dbSNP's Web Site and XML Files

Using e-utilities to fetch dbSNP entries in XML format, I find that the <Rs_Sequence_Observed> for rs2066844 is A/C/T/G, but if I get rs2066844 from the dbSNP web site, the alleles are shown as C/T.

The dbSNP RefSNP page and efetch should provide the same data.

In this particular case, however, a submitted SNP(ss) that contributed to the extra alleles was withdrawn (probably due to strand or sequence error), but there wasn't time to make corresponding change in the annotation before the build was released. The annotation will be fixed in the next build (b127), which should be released very shortly. (**12/18/06**)

The dbSNP summary page says there are 43833 new RefSNPs for build 127, but the XML files I downloaded this week have no RefSNP clusters with a "create build" equal to 127.

The new refSNP cluster number is determined by comparing the refSNP cluster creation date with the last build (b126) release date (which was May 4, 2006).

Here is example XML for rs41318644:

```
<Rs>
<Rs_rsId>41318644</Rs_rsId>
<Rs_create>
<Rs_create_build>126</Rs_create_build>
<Rs_create_date>2006-11-29 15:30</Rs_create_date>
</Rs_create>
```

So you are correct in thinking that the rs_create_build_id should be b127. This build_id assignment rule will be fixed for the next build which we plan to release sometime in July. In the meantime, if you are only interested in new refSNPs, could you use the <Rs_create_date> by comparing it with "May 4, 2006"?

Please also note that in b127, all refSNPs (both old and new) have been mapped to the new genome map (NCBI build 36.2) and so new gene annotations are based on this new map. (**5/25/07**)

dbSNP shows that there are 27 million human submitted SNPs (ss) in build 125. The xml files we obtained from your ftp site, however, show only about 21.8 million submitted SNPs for human.

The XML file report for the refSNPs lists only submitted SNPs (ss) that occur within refSNP clusters. At the release of build 125, there were ~5 million newly submitted SNPs that did not get into the clustering pipeline in time for the build release, and were not recognized as clustered. Hence, these submitted SNPs did not make it into the XML report. The ~5 million unclustered submitted SNPs have now been clustered and will be released in the next human build.(**4/26/06**)

I can find genotypes and allele frequencies for the ss4970601 population “HapMap-CEPH-30-trios” on the dbSNP web site for the new build but cannot find this information anywhere in your XML files.

The GenoExchange FTP file has not been created for the new build. If you have a specific set of SNPs you are interested in, you can create an XML file with information from the new build by using the Entrez SNP batch query (i.e., dbSNP batch report) and the Genotype report format option.

Data Formats

Why does dbSNP have two different XML formats (one for the file and one for data for EUtils) for SNP entries?

We have plans to merge the two XML schemas in the near future. As for the schema merge it will be done some time this year. As we have a lot of ongoing projects currently, I’m unable to be specific about when this will be accomplished.(**2/5/07**)

Decompressing XML Files

I can unzip all of the other zipped SNP XML human chromosome files, so why can’t I unzip the SNP XML file of human chromosome 1 after downloading?

It is possible to unzip this file, but because it is so large (the unzipped file is 4.3 Gb or 4393989049 bytes), you must use a UNIX machine, or try PKZIP or WinRAR archivers. WinZip does not handle files over 4 Gb. In the meantime, we will split files over 4 Gb into smaller files before we zip them.

Functional Data

In the dbSNP XML files, where do I find the number of SNPs in coding vs. non-coding regions?

The gene information is encoded in the XML for those genes where we are able to map SNP(s) in your organism of interest. docsum_2005 to determine whether the SNP is coding. You can get the coding counts by grouping the coding-synon | coding-nonsynon | reference | together. You can get the non-coding counts by grouping the mrna-utr | intron |splice-site | locus-region | together.

You can also obtain these counts for all genes in the human genome using Entrez SNP. Select the Preview/ Index function and then set the Function Class limits at the top of the form, under the Limits link.

When I used the last method, Entrez SNP produced the following result:

#11 Search human[ORGN]
Limits: intron, locus region, mrna utr, splice site
COUNT=1998902
#10 Search human[ORGN]
Limits: coding nonsynon, reference, coding synonymous
COUNT=49522

Allele Data

I'm confused about the representation of some of the alleles in the XML files. In the dbSNP XML files, I have found some strange alleles (+, +/T, D, N, NNN, etc.), and that separate alleles are described in some cases, while wobble codes are used in other cases. Why is this?

dbSNP allows many "bad alleles", because our parser rules include the following:

1. Accept all IUPAC codes and a "-".
2. If an allele is within parentheses, then it is treated as a description of an allele of the variation when a submitter does not know the actual allele, or when the allele bases are longer than 250.
3. Do not allow common bases in the beginning or ending of all the alleles of a variation. For example, the "NNN" allele was used in variation "NNN/-". This means a three-base insertion, but the submitter did not know the actual bases in it.
4. C/Y means a submitter clearly detected a "C" in some sequences, whereas in other sequences, the base detection was indeterminate—the submitter knows that there might be a "C" or a "T".

We still have some bad alleles that I am in the process of cleaning up, including the two in your list ("+", "+ T"). There is also a "+G" in variation "+G/-". These bad alleles involve four SNPs. I will email the submitters for clarification and not allow this type of allele form in future submissions.

I'm using the data exchange format of dbSNP's XML files and noticed that there are two allele values ("N" and "+") that I haven't seen before. What do these values represent?

"N" has two meanings, depending on the context in which it is used. The first context in which "N" is used is allele frequency. In this context, "N" means "indeterminate frequency". For example, if a submitter has a sample size of 120 chromosomes, they may submit A=40/C=78/N=2.

The second context in which "N" is used is in SNP FASTA sequence. Here, a variation is represented by the IUPAC letters of A, C, M, G, R, S, V, T, W, Y, H, K, D, B, and N. If the variation is not represented by one of the first 14 letters, then it is considered an "N". For example: All indels, microsatellites, and named variations are expressed as "N" in SNP FASTA sequences.

Some submitters in the past have used "+" to represent the insertion part of an indel SNP. You could get the real inserted sequence (relative to the deletion) from the variation from the SNP assay. We realize that allowing "+" may confuse users, so we are in the process of substituting the real "insertion" sequence for the "+" currently available and hope to get this done by build 122.

Genotype Data

How do I upload the data for only those SNPs that have genotype or frequency information?

There are two ways to upload the data for all SNPs that have genotype or frequency information.

The easiest way to do this is to upload the Genotype XML GenoExchange files. These files contain all dbSNP Genotype and allele frequency information arranged by chromosome for each organism. For example, the human build 125 genotype data are located in the genotype subdirectory of the human SNP directory. The schema for the GenoExchange files is also located online.

Another way to upload the data for all SNPs that have genotype or frequency information that requires a little more work, is to upload all the dbSNP database files, which are also available on our FTP site. (**5/6/06**)

Are the genotype data available as XML files different from those available in .bcp-formatted files, or are they just organized or formatted differently?

The XML files are generated from the database from which the .bcp files are created. There should be no difference between them, apart from some minor formatting. Please let us know if you detect anything unusual.

I have just downloaded the b125 XML files from the dbSNP ftp site and can't find the population/frequency information. Why did you take it out?

Population and frequency information is now located in the genotype files for the organism of interest. You can find this file by going to the dbSNP ftp organism directory, select an organism of interest (in this case I will choose "human_9606"), and then select "genotype" from the list of the organism's subdirectories. **(11/4/05)**

I downloaded human XML and ASN reports for build 125, but found that many of the SNPs in these reports do not have population frequency data.

Some submitters did not submit genotype or frequency data to dbSNP in their submissions; therefore, there is no population frequency data for these SNPs. There are approximately 27 million submitted SNPs in dbSNP, and only 3.5 million of those have frequency data associated with them. (**1/9/06**)

Mapping Data

SNP rs4247888 maps to two positions, but the dbSNP XML files show this SNP in ds_ch4.xml and not in ds_chMulti.xml. What am I missing?

If a SNP hits once or twice on the same chromosome, it is assigned to a chromosome file; if it hits more than twice or hits on different chromosomes, it goes to ds_chMulti.xml. This was designed to take account of possible fragment redundancy in unfinished parts of the genomic sequence. There is one notable exception, however. When SNPs hit in the pseudo-autosomal region on Y, they are recorded on both the X and Y chromosome files. Also, we track hits to both the reference genome as well as the alternate assemblies and haplotypes. When a SNP hits on several alternate scaffolds, we record it several times but consider only the number of distinct loci when deciding whether to assign it to chMulti.

I'm looking for a SNP that seems to have no genome mapping positions. I thought it would be in the 126 XML ds_chUn or ds_chNotOn files, but can't find it. Where is it?

The SNP you are looking for is in the ds_ch11.xml file of the human XML directory. This SNP maps to the Celera assembly only, so it will not appear in your Entrez search results since Entrez indexes only SNPs that map to the NCBI reference assembly.(**8/30/06**)

Retrieving a XML Report for a Specific Submitted SNP (ss)

How do I retrieve a specific xml record for a submitted SNP(ss242) in a format that is similar to “rs_ch1.fas”, which contains the sequence information for rs242?

You can use eUtils or dbSNP Batch Query to retrieve to retrieve specific xml record(s) containing submitted SNP (ss) information. You might want to take a look at this eUtils example for rs242, as well as online eUtils tutorial and examples . **(12/01/05)**

Retrieving XML Data for an rsID

How do I fetch data for a given refSNP ID (rsID) in XML format using a simple URL-request?

Please see the online short course for using eutils (**9/15/06**)

Sequence Data

The refSNP page for rs28928880 shows the refSNP’s amino acid position to be 25, but b126_SNPContigLocusId_36_1.bcp file for human shows the amino acid position of rs28928880 as 24.Why are these data different?

The sequence coordinate data for the XML, ASN.1, .bcp, and the Genotype/genotype_by_gene files were changed from 1-based to 0-based starting with dbSNP build 125. The ASN.1_flat, Chromosome Report, and the web page reports remain 1- based. (**6/30/06**)

Varview Data

I’ve noticed that since the release of build 129, I’ve noticed “VarView”, and was wondering if the information it contains will be available in flat files.

We are glad that you are interested in the new VarView. It is still in a very early stage of release. We will be working on providing XML dumps for VarView data in the future. (**05/08/08**)

Tags/Flags

Where can I find definitions for the various tags/flags in dbSNP’s XML?

The element FlagDesc in the xml lists the flags and their descriptions. **(4/8/05)**

What is the relationship between dbSNP’s XML tags and the tables and fields in dbSNP's relational schema?

The correspondence between the XML tags and the dbSNP schema tables is as follows (the “vw” prefix means that it is a view rather than a table):

XML tag	Schema Table
<NSE-rs_refsnp-id>	SNP.snp_id
<NSE-rs_create-build>	vwSNP_build.create_build_id
<NSE-rs_update-build>	vwSNP_build.last_updated_build_id
<NSE-rs_observed>	ObsVariation.pattern
<NSE-rs_seq-5_E>	SubSNPSeq5.line (set of rows, ordered bySubSNPSeq5.line_num)
<NSE-rs_seq-3_E>	SubSNpSeq3.line (set of rows, orderedbySubSNPSeq5.line_num)
<NSE-rs_het>	SNP.avg_heterozygosity
<NSE-rsMaploc_asn-from>	SNPContigLoc.asn_from
<NSE-rsMaploc_asn-to>	SNPContigLoc.asn_to
<NSE-rsMaploc_physmap-str>	SNPContigLoc.phys_pos

```
<NSE-rsMaploc_physmap-int>              SNPContigLoc.phys_pos_from
<NSE-rsContigHit_accession>             SNPContigLoc.contig_acc
<NSE-rsContigHit_version>               SNPContigLoc.contig_ver
<NSE-rsContigHit_chromosome>            SNPContigLoc.contig_chr
<NSE-FxnSet_symbol>                     SNP.ContigLocusId.locus_symbol
<NSE-ExchangeSet_dbSNP-build-number>    Not in the database, set externally
```

Can you please clarify your use of the "INTERIM" Gene Symbol?

dbSNP propagates the INTERIM tag from Entrez Gene. Although the INTERIM tag is not unique, the locus_id for each gene is distinct and is associated with an mRNA transcipt and protein in the NCBI refseq database (and sometimes more than one, in the case of splice variants).

The INTERIM tag is assigned during the human genome annotation process. Think of INTERIM as a flag indicating that genes have been predicted at this locus, based on mRNA and protein models, but have not yet been curated to a known gene symbol.

FTP File Descriptions and Definitions

Can you explain the difference between a refSNP that is mapped to a chromosome but is "unplaced", versus a refSNP in Chr_Un? What is Chr_Un?

"Placed" contigs are contigs with clear starting and ending positions on a chromosome.

A contig is called "unplaced" if it meets one of the descriptions below:

- If the contig is known to be in a chromosome, but the contig position in the chromosome is unknown.
- If the contig's chromosome is not known.

NT_113953 is an example of an "Unplaced contig".

Other examples of "unplaced' contigs include the following:

- In build 36.3, the NCBI reference assembly has 8 contigs that are known to be in a chromosome but have unknown chromosome positions
- HuRef has 3110 contigs with unknown chromosome placement
- Celera has 9 contigs with known chromosome but unknown positions and 5898 contigs with unknown chromosome origins

"chr_Un" is a file name for a file in which we place SNPs (with the exception of weight 1 and weight 2 SNPs) that map to "unplaced" contigs.

Please see the FAQ that explains mapping weight.

The reason that weight 1 and weight 2 SNPs are not placed in "chr_Un" is that when a SNP maps uniquely to a placed contig, but also maps to an "unplaced contig", we ignore the "unplaced contig" placement when we assign mapping weight. So it is as if the SNP only maps uniquely to the "placed" contig. (**07/14/08**)

Where can I find the specs for SNP_bitfield.bcp.gz?

The specs for SNP_bitfield.bcp.gz are located in the specs directory of the dbSNP FTP site. (**11/08/07**)

The Entrez directory in the dbSNP FTP site contains a file called snp_omimvar.txt. What is it used for? What does it contain?

snp_omimvar.txt is used by OMIM to indicate which allelic variants have been matched to refSNP (rs) numbers.

You can find a description of this file's use and content in the snp_omimvar.README, which is also located in SNP's Entrez directory. It states that snp_omimvar is a tab-delimited file that contains the following columns:

1. Identifiers in dbSNP (rs number, as snp_id)
2. MIM number of the record that has links to dbSNP
3. Allelic variant number of the variant that corresponds to the record in dbSNP. (If this value is 0000, the link is NOT based on an allelic variant).

This file is updated every Monday, and the method of computation is as follows:

Connections between an rs number and MIM numbers result from the rs number being reported in the text of an OMIM record.

Connections between an rs number and an OMIM allelic variant are established when:

1. An rs number is within the text of the description of an allelic variant
2. A variant is mapped to sequence and thus to an rs number
3. A submission indicates that an allelic variant corresponds to an rs number.

(**8/30/07**)

Where can I find a description for chr_rpts? I want to know which column in chr_rpts represents the contig orientation of a SNP hit.

The description of the chr_rpts files are in the dbSNP FTP readme file.

The Chr_rpt column definitions are located about three-quarters of the way down the FTP readme file under the "CHROMOSOME REPORTS" section heading. Although SNP orientation is not reported in the chr_rpts files, you can find SNP orientation by looking at the entry for a specific refSNP(rs) number in the ASN1_flat files. Look for SNP orientation in the CTG line of the entry for the rs number of interest. Below is an entry for an rs number taken from the ASN_flat files:

```
rs8896|human|9606|snp|genotype=NO|submitterlink=YES|
|updated 2004-10-04 13:37|ss10932|CGAP-GAI|52782|orient=+|
|ss_pick=YES SNP|alleles='C/T'|het=?|se(het)=? VAL|validated=NO|
|min_prob=?|max_prob=?|notwithdrawn CTG|assembly=reference|
|chr=MT|chr-pos=8270|NC_001807.4|
|ctg-start=8270|ctg-end=8270|loctype=2|orient=
```

(**2/14/06**)

Where can I find documentation for the genotype data exchange format?

The genotype data exchange (genoex) format documentation (schema) is available online. (**5/8/06**)

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)**

Where can I find definitions for the various tags/flags in dbSNP's XML?

The element FlagDesc in the xml lists the flags and their descriptions. **(4/8/05)**

Field Definitions

What are the definitions of contig_chr, contig_pos, and contig_build_id?

contig_chr is the chromosome for this SNP (if known).
contig_pos is the position on the chromosome.
contig_build_id is the build in which the SNP was last edited.

What is the definition of "contig_pos=0"?

This is a case in which the SNP has not been mapped to the human genome reference sequence.

Specific File Location

The ftp download site contains the data file b127_SNPChrPosOnRef_36_2.bcp.gz. Do you have a corresponding data file for build 35?

b127_SNPChrPosOnRef_36_2 is a new table starting b127. We do not have the same data file for genome build 35. (**5/17/07**)

In build 124, there is a table called "SnpFunctionCode.bcp.gz", but I'm unable to find this table in the organism specific directory.

This table is shared among organism databases, so it is located in the shared_data directory of the dbSNP FTP site. (**1/18/06**)

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**)

Report Formats Not Available on dbSNP's FTP Site

Can you give me advice on downloading dbSNP in a format that can be opened by Microsoft Access?

We do not have output in Access database format. Please take a look at our ftp site where we have a tab-delimited table dump that could be loaded into Access. I should mention that directly importing the data to Access is not a good idea since some of these tab-delimited tables are quite large (> 100+ Mb). Alternatively, you could get your data in alternate formats (like XML) that we have available on our ftp site if you can narrow down the number of SNPs you will be using for your research. **(2/23/05)**

Is it possible to obtain the SNP genotyping data in Excel format?

It is possible, but the XML genotype reports have a number of nested element dimensions that are difficult to convert into the two-dimensional, row-and-column format of an Excel worksheet. We have a tool to do the conversion, but Microsoft discusses a method for the import of XML into Excel using an XSLT template.

How do I get a tab-delimited report of all mouse SNPs from dbSNP that show refSNP ID and fxn_class?

Custom reports are not available. You'll have to parse the data from the XML files on the FTP site.

Why has the support for the SQL server backups been removed?

NCBI has a long-standing policy to avoid any use of vendor-specific file formats on our FTP site. We provide public data in a generic format so all users have the same degree of access and utility, regardless of their specific platform or technology. NCBI, however, cannot be placed in the position of endorsing any specific vendor's format or technology. We are a government agency, and such an act is illegal. Seemingly simple acts such as posting vendor-specific backup files can carry the weight of an endorsement.

Given our legal obligations to treat all vendors fairly, the decision to redistribute dbSNP in a generic table/schema format is unlikely to change. The best we can do is work with users to ease the use of these generic files. We have posted some sample scripts for loading and verifying data.

Do you have any files on your FTP site where the amino acid changes are written in this format: E429A ?

We currently do not have amino acid changes written in the format: "E429A". I know that a member of the SNPdev group is working on updating the database to allow searches for SNPs by amino acid position when search term is entered in the "E429A" format.

It is possible that in the future, we may have this format for FTP download. As it is hard to determine a timeframe for moving the "E429A" format to FTP right now, you may wish to subscribe to dbSNP announcements to keep abreast of the latest updates to dbSNP.

Until your format becomes available, you can find amino acid positions and changes for SNPs located within a gene on the “GeneView” page. Review an example showing all the SNPs linked to the gene LPL. Here you can see that the amino acid change for rs11570895, using your format, would be V28A. We count the amino acid position from the start codon ATG. (**2/23/07**)