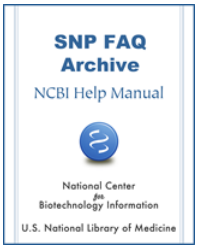
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: SNP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. Submission Formatting. 2005 Jun 6 [Updated 2010 Jun 15].
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Submission Formatting

Created: June 6, 2005; Updated: June 15, 2010.

General Format Information for Submission

What format do we use to submit to dbSNP?

Submissions can be sent as plain text or as word docs, but dbSNP has an excel template that might make the submission process a bit easier.

Does NCBI have any scripts for file format validation that will check for proper formatting, required fields, etc?

We do not have a script that will do what you requested. All file format validation is completed while loading to dbSNP.

(**11/06/06**)

Do I send SNP submission data as a text document or key all of the data into an email?

You can send your submission as a text file. If you send us a word file, we must first convert it to text before loading the data. **(3/11/05)**

Do I have to submit all my SNPs individually, or I can put them all in one large file?

You can submit all your SNPs using one large file. Just repeat the SNP section under the SNPASSAY header. Place the appropriate SNP type in the comment section and repeat this section for each SNP.

I was wondering if NCBI has some universal recommendations concerning the reporting of new SNPs for our publication. Should we use HGVS standards?

I don't know specific NCBI policy on offering "universal recommendations" to publishers about their citation policies, but I will point out that dbSNP identifiers [submitted SNP (ss) number; refSNP (rs) number] are stable and unique identifiers within the NCBI dbSNP database, and that they provide flanking sequence context and alleles for a specific variation.

The issue of requiring use of HGVS nomenclature as you have asked is useful when all the information pieces are available (i.e. accessioned sequence records, gene annotation, functional analysis). Many polymorphism discovery projects, however, occur before such organized knowledge is available for a new species genome, and the rs number/ss number paradigm has been sufficient to describe sequence variations at this minimal level of detail.

I would say use HGVS standards when available, but use the rs numbers and ss numbers as well. This will encourage submission of the variation data to a public repository since using HGVS alone does not guarantee that the variation is in a public database. (**8/14/07**)

Are you going to accept submissions of new SNPs in an XML format (e.g. /genoex_1_4.xsd DTD) anytime in the future?

Currently we don't have anyone in dbSNP to work on providing xml submission support, but if you want to submit genotype data as xml it should be fairly easy for us to transform using xslt to dbSNP flat text format.

We would also need to capture the meta data (contact, pub, method, etc.) associated with the genotype data for which there is a schema that we can reuse. (**06/23/08**)

Sequence Formatting

What do I do when submitting SNPs from short sequence reads (Solexa or 454) since these reads might not contain the required 25 base pairs of flanking sequence?

dbSNP requires at least 25 bp of flanking sequences in order to cluster and map the SNP to the genome or reference sequences. You can extend your SNP flanks to the required 25bp using contig sequence.

When you submit, place the short reads (<25bp) flanking the variation in the 5'_ASSAY and 3'_ASSAY fields and the extended sequences in the 5'_FLANK and 3'_FLANK fields. (**07/02/08**)

How do I indicate genetic distance in a submission?

To my knowledge, we do not take genetic map (genetic distance) information for dbSNP submissions. The minimum information required for a submission is a variation's observed alleles, the 5′ and 3′ flanking sequences, the name of the gene in which the SNP is located, and the NCBI GenBank accession number of the genomic sequence in which the SNP is located. (**8/27/07**)

How do I submit a sequence that contains multiple SNPs, all of which are linked? Do I submit them together or should I submit each SNP alone?

Please submit each SNP individually regardless of whether or not they are linked. If adjacent SNPs are part of the flanking sequence of the SNP you are submitting, you can put an "N" at that position, or other more specific IUPAC codes. (**09/05/07**)

I am submitting a SNP whose flanking sequence contains variations. What notation do I use for variations in the flanking sequence?

Use the IUPAC code for degeneration. (**10/8/07**)

I'm submitting SNPs identified using minisequencing, so have only the 5' adjacent region of the target SNP, and a 400 bp PCR product. Problem is, both the "5'_flank" and "3'_flank" fields are mandatory if the assayed sequence is less than 25bp.

The flanking sequences that you submit will only be used to locate the genomic position of a SNP, so as long as you know the sequence, it does not matter how you got them. The total length of 5'_Flank + (5'_assay) + observed + (3'_assay) + 3'_flank must be at least 100 bp, with each component being at least 25 bp (except the "observed" component of course), and the assay sequences are optional. You can simply provide 100 bp of the sequence. In your case, since you can amplify a 400 bp region by PCR, you should have no problem finding the 100 bp sequence you need. (**09/04/07**)

In a dbSNP submission, if I have SNPs contained within the assay and/or the flanking sequence do I represent them as seen in the IUPAC code?

Yes, you can use the IUPAC code. (**5/24/07**)

I know we can use the IUPAC ambiguity code to indicate SNPs in the flanking regions of a submission, but how do I indicate the presence of indels in the flanking regions?

You can use "N" to represent an indel SNP. (**08/17/07**)

Can I format the variations in sequences I put into the 5'_ASSAY and 3'_ASSAY sections of my submission form using something like {C/A}?

The format you suggest {C/A} is not allowed in the flanking sequences. Use the IUPAC Ambiguity Codes to code for variations in flanking sequence. Remember, the flanks should each be at least 25bp, and their sum should be at least 100bp. (**3/21/07**)

We want to submit computationally derived SNPs to dbSNP, but need to know if strand orientation matters. We have SNPs on both the positive and negative strands.

Either strand is fine as long as a SNP allele is reported in the same orientation as its submitted flanking sequence.(**5/3/06**)

A number of SNPs that I intend to submit are very close to each other on the sequence. How do I format these SNPs for the 5'FLANK, OBSERVED, and 3'FLANK portions of the submission?

Use IUPAC ambiguity characters for a known SNP that is contained within the flanking sequence.

(**5/17/06**)

Can we include characters like "Y" and "M", etc., in the flanking sequence for the submission?

Yes, you may use the IUPAC codes (Y, M, etc.) in submitted flanking sequence. (**2/17/06**)

How do I submit a pattern like this: 5' atttcaaat/gaatcggggc 3' (t/g is the polymorphic site)?

Break the sequence up into the 5'flanking sequence, the 3' flanking sequence, and the observed variation. Then place the sequences after the appropriate tags in the excel submission file as shown below. Remember to provide at least 100bp of flanking sequence.

```
5'_FLANK: atttcaaa
OBSERVED: t/g
3'_FLANK: aatcggggc
```

(12/14/05)

Allele Formatting

How do I indicate genetic distance in a submission?

To my knowledge, we do not take genetic map (genetic distance) information for dbSNP submissions. The minimum information required for a submission is a variation's observed alleles, the 5′ and 3′ flanking sequences, the name of the gene in which the SNP is located, and the NCBI GenBank accession number of the genomic sequence in which the SNP is located. (**8/27/07**)

SNPs are usually, but not always heterozygous (i.e. A/AG). How do I annotate this in a submission: A/G or A/AG?

Annotate "A/AG" as "-/G". The common allele "A" should not be included in the OBSERVED line. **(12/01/05)**

How do I report an AA deletion, and the insertion of GC in the place of the AA (or AA>GC)? It was reported as an insertion/deletion in a NEJM paper because of nomenclature requirements.

This is a multiple nucleotides polymorphism (MNP). The variation should be reported as AA/GC on the "OBSERVED:" line in the submission file. **(5/4/05)**

Formatting Submissions of Specific Data Types

We have identified novel in some cell lines. Since this is an in vitro system I am not sure whether these variations can be submitted to dbSNP

dbSNP does accept data derived from an in vitro system; just specify in your submission that the SNP is known to be a somatic variation by adding the following line in the SNP section:

```
SOMATICYES
#
```

You can comment about the variation in the line that occurs after the "#" if you wish. Such a comment is optional. (**01/10/08**)

We want to submit SNPs, their flanking sequences, and their allele information that were obtained *in silico*. How do we submit these data? Do we also need to submit to dbSTS?

Use the SNPASSAY format. Many fields in the SNP section are optional; you just need to provide the following required fields: local snp id, GenBank accession, and allele 5_flank. Because your sequence is *in silico*, use 5_FLANK, as well as 3_FLANK, 5_assay. 3_assay is the sequence obtained from a laboratory method.

With regard to STS submissions, do you have new STS primers that are not in dbSTS? If you do, fill in as much as you can in the STS sections. We don't parse the STS sections. We just pass the information through to dbSTS. The dbSTS administrator will contact you if more information is needed. If you are using existing primers in dbSTS, provide the dbSTS accession in the SNP section tag STS. If you don't have STS information, you won't need to provide the STS sections and won't have to provide STS accession.

The best way to proceed is to prepare a submission and send it to us. We will test load it, report any problems, and send suggestions back to you.

I have used in silico methods to predict SNPs from EST data, and have produced validation criteria for screening out false positives. How do we submit fill out the submission form for these SNPs?

The only sections in the submission form you need to fillout for in silico SNPs are CONTACT, PUB (optional), METHOD, and SNPASSAY header and SNP(s). You can skip the other sections.(**8/23/06**)

Does dbSNP use TOP/BOT nomenclature, and can I use the Top/Bottom strand designation information in a frequency submission?

dbSNP does use Top/Bottom strand designation information for submitted SNPs when that information is easily obtained. However, we do not use the Top/Bottom strand designation information in frequency submissions for the following reasons:

1 The Top/Bottom strand designation rules cover four distinct situations. The first two situations (A/G and A/C) are unambiguous so the Top/Bottom strand designation rules work well. When alleles are symmetric (A/T and C/G), however, the strand designation rules rely on flanking bases at equal distances to variation site being non-symmetrical. Relying on neighboring base symmetry poses the following problems:

After having scanned the whole of dbSNP, I found a small but significant percentage of SNPs did not meet this condition (flanking bases at equal distances to variation site being non-symmetrical), so the Top/Bottom strand could not be defined.

dbSNP takes frequency submissions for variation classes (indels, multiple case substitutions, and microsatelites), that have no rules for the Top/Bottom strand designation. For symmetrical alleles in these classes, I have tried using the same rule (flanking bases at equal distances to variation site being non-symmetrical) to determine Top/Bottom strand designation and again found that a small but significant percentage of SNPs did not meet this condition, so the Top/Bottom strand designation could not be defined.

Some genomic regions contain SNPs so densely packed that flanking sequences also contain variations; in such a case, the Top/Bottom strand designation would not be stable.

1 Although most of dbSNPs variations have two alleles, dbSNP also includes variations that have 3 or 4 alleles — such as SNPs in the highly variant HLA regions.

Although it will be difficult to use our strand code for a frequency submission, we'd be glad to discuss how the code could be applied to your frequency submission files if you contact us at snp-admin@ncbi.nlm.nih.gov. (**8/29/06**)

We are submitting a mixed population of cattle SNPs. How do we indicate that the only heterozygotes observed in this mixed population were in the *Bos taurus* × *Bos indicus* F_1 crosses?

dbSNP submissions are organized by species and/or species cross. In your case of differences between *Bos taurus* and *Bos indicus*, there are three possible NCBI taxonomy IDs that can be used to segregate the data by organism.

Data from a Bt × Bt cross should be assigned to taxid 9913, and Bt × Bi cross data should be assigned to taxid 30523. Please note that the POPULATION section of the dbSNP submission is a free-text data field, where you can label each mating type and describe any pertinent details of sampling and cross-design.

–

Cross	Taxonomy ID	GenBank Common name	Rank
B. taurus x B. taurus	9913	cow	species
B. taurus x B. indicus	30523	hybrid cattle	species
B. indicus x B. indicus	9915	zebu	species

The following set of node descriptors from the NCBI taxonomy database illustrate specificity at a variety of taxonomic ranks including Genus, species, sub species and hybrid cross.

Bos (oxen, cattle) Click on organism name to get more information.

Bos taurus (cow)

Bos indicus (zebu)

Bos taurus x Bos indicus

Bos indicus x Bos taurus (hybrid cattle)

Bos indicus gudali (Gudali zebu)

Bos javanicus (banteng)

Bos primigenius (aurochs)

Bos sauveli (kouprey)

Bos frontalis (gayal)

Bos gaurus (gaur)

Bos grunniens (domestic yak)

Bos grunniens mutus

Bos sp.

A sample set of population sections might look like:

```
TYPE: POPULATION
HANDLE: <yourhandleID>

ID: USDA Bos taurus

POPULATION: Samples were collected from two outbred populations of B. taurus stock in
Texas, USA and Alberta, Canada.

TYPE: POPULATION
HANDLE: <yourhandleID>

ID: USDA Bos taurus x Bos indicus cross

POPULATION: B. taurus samples were collected from two outbred populations of B. taurus
stock in Texas, USA and Alberta, Canada. B. indicus samples were collected from a
USDA-managed population in western Maryland. F1 progeny assayed for nucleotide variation
were derived from B.t. males and B.i. females.
```

In the snpassay and individual genotype section(s), these populations are referenced

as appropriate:

```
TYPE: SNPASSAY
HANDLE: <yourhandleID>
BATCH: 4-1-2004-1
MOLTYPE: Genomic
METHOD: <your method ID>
SAMPLESIZE: <# chromosomes assayed for variation>
ORGANISM: Bos taurus x Bos indicus
POPULATION: USDA Bos taurus
||
SNP: mysnp1
LENGTH: 100
5'_ASSAY: CTTCGTTCCTACACACTTTC
OBSERVED: C/T
3'_ASSAY: AGCCTGCCTACCCTA
||
SNP: mysnp2
||

TYPE: SNPASSAY
HANDLE: <yourhandleID>
BATCH: 4-1-2004-2
MOLTYPE: Genomic
METHOD: <your method ID>
SAMPLESIZE: <# chromosomes assayed for variation>
ORGANISM: Bos taurus
POPULATION: USDA Bos taurus x Bos indicus cross
||
SNP: mysnp3
LENGTH: 100
5'_ASSAY: CCTGAGAGCCTACCCTACTTA
OBSERVED: G/T
3'_ASSAY: GGGCTTTTTCCCACCACAC
||
```

```
SNP: mysnp4
||
```