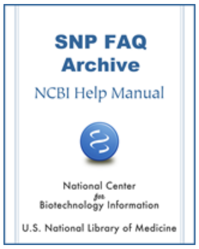
U.S. National Library of Medicine
National Center for Biotechnology Information

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


SNP FAQ Archive
NCBI Help Manual
National Center for Biotechnology Information
U.S. National Library of Medicine

Filling out Specific Sections/Fields within a Submission Worksheet

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

The ACCESSION field

What do I put in the "ACCESSION" submission field?

In the ACCESSION filed you would put a GenBank accession number that represents the sequence you are submitting. **(7/29/05)**

Where can I find a GenBank accession number for the reference sequence I need to place in the submission “ACCESSION” field?

Go to Entrez Nucleotide to get the GenBank accession number.

If you don't know the accession number, then leave the ACCESSION field blank. **(12/14/05)**

If I submit a SNP whose underlying sequence has already been submitted to GenBank, can I use a direct link to the SNP sequence, or is a specific kind of link needed?

If you have submitted the underlying sequence that defines the SNP to GenBank, then use the GenBank assigned accession number. Otherwise, you can omit the ACCESSION field.

When I submit, how do I reference BAC endclone sequences that don’t have accession numbers, and since I did not generate these sequences, I cannot submit them to Genbank?

You can use a different ID to reference these sequences in your submission files. Just put a description of the ID in the “Batch Comment” section of the submission file. It would also be nice if you could include the URL for each of the sequences, if they are available. If you are still having difficulties, please submit the data to dbSNP, and we will help you with the formatting.(**10/16/06**)

The SNP_ID field

What are local SNP identifiers?

These are identifiers used by a submitter to identify a submitted SNP, a population, an individual within a population, a method for assaying SNPs, or a set of submitted SNP assays/experiments. Once the data with their local identifiers has been submitted, dbSNP will assign standardized identification to the data, but the submitted data record will retain this submitter-created identifier (an identifier that is “local” to the lab which generated the data). Users can search dbSNP for submitted data in dbSNP by using either the dbSNP assigned standardized identifiers, or by using the HANDLE | LOCAL SNP ID combination.

Local identifiers need only be unique within the data submitted under a specific dbSNP handle, so the combination "HANDLE | LOCAL SNP ID" will be unique within dbSNP. Please note that there is a 64 character limit for local SNP identifiers. Further information on identifiers and examples of identifiers is available online. (**04/05/06**)

How do I create submitted SNP (ss) and RefSNP (rs) numbers for my submission?

dbSNP assigns the ss and rs numbers after your submission has been loaded (ss) and clustered (rs). In the meantime, you can use your local SNP ID/name instead of an ss number or rs number when reporting genotype (SNPINDUSE) or frequency data (SNPPOPUSE). **(5/16/05)**

The POPULATION section

I would like to submit allele frequency information by ethnic group. If we have four ethnic groups, do we create four POPULATION entries?

Correct. (**03/21/08**)

If the POPULATION section applies to all the SNPs in a submission, do we need to include it before each SNP entry?

No, the POPULATION description section only has to appear once. Place it at the top of the submission.

I used subjects from a number of populations housed in the Human Diversity Collection, Coriell Cell Repository, in a study and want to submit the data to dbSNP. How do I address subject IDs in my submission?

You can use the individual or sample IDs provided by the source (i.e. CEPH, Coriell, etc.). Please see the "Individual Description Section" of the online dbSNP submission documentation (**9/19/06**)

The POP: ID field

I am submitting a number of plates that contain multiple tumor types, and want to know how I should fill out the POP_ID section, and where I should report frequency data.

The POP_ID field is for designating the geographic origin of the population. You can use the value "UNKNOWN" in your case unless you know the geographic origin of the tumor donor. Please see online documentation for population descriptions.

The frequency data should be reported in the SNPPOPUSE section. Please see online documentation for submitting population variation information. **(12/01/05)**

When I submit original data, do I shorten POPULATION_ID to POP_ID and follow it with POPULATION, whether or not I have data for the POPULATION field?

Correct. The POPULATION tag is required whether a description is included or not.

In the submission form, how do I go about naming the method ID, the population ID, and the batch ID? Are there specific naming rules?

The method ID, the population ID, and the batch ID are also known as a local identifiers (an ID that is "local" to your lab). You can assign any name/ID or local identifier you can think of, but it must be unique for each method/population/batch you submit. Avoid using "||", "|", quotation marks (" ") or a colon (:) when naming, and note that there is a 64 character limit for this ID.

The Submission templates available on the dbSNP FTP site now include instructions for each field. You can also go to the dbSNP submission Quick Start for more information about filling out your submission form. (**5/22/07**)

The METHOD_CLASS field

We used DHPLC and sequencing to identify the SNP we want to submit, but used pyrosequencing for assaying the SNP. Which method class should we use in the submission form?

Valid classes for the METHOD_ID field include the following: Sequence, DHPLC, Hybridization, Computation, SSCP, Other, Unknown.

For the SNP Assay field, you can use DHPLC, or you can choose another assay type — what you put in this not critical, what is however, is that you put a clear description of your method in the method field. (**2/14/07**)

The Method: ID field

In the submission form, how do I go about naming the method ID, the population ID, and the batch ID? Are there specific naming rules?

The method ID, the population ID, and the batch ID are also known as a local identifiers (an ID that is "local" to your lab). You can assign any name/ID or local identifier you can think of, but it must be unique for each method/population/batch you submit. Avoid using "||", "|", quotation marks (" ") or a colon (:) when naming, and note that there is a 64 character limit for this ID.

The Submission templates available on the dbSNP FTP site now include instructions for each field. You can also go to the dbSNP submission Quick Start for more information about filling out your submission form. (**5/22/07**)

The BATCH field

What is the definition of the "Batch" submission field?

This field is for a user-defined Batch Id associated with your submission. You can name it whatever you want; some people use dates, others use lab-specific IDs. **(5/16/05)**

In the submission form, how do I go about naming the method ID, the population ID, and the batch ID? Are there specific naming rules?

The method ID, the population ID, and the batch ID are also known as a local identifiers (an ID that is "local" to your lab). You can assign any name/ID or local identifier you can think of, but it must be unique for each method/population/batch you submit. Avoid using "||", "|", quotation marks (" ") or a colon (:) when naming, and note that there is a 64 character limit for this ID.

The Submission templates available on the dbSNP FTP site now include instructions for each field. You can also go to the dbSNP submission Quick Start for more information about filling out your submission form. (**5/22/07**)

The PUBLICATION section

If I submit to you before I submit to a publisher, do I leave the publication section of the SNP submission form blank?

You can either leave it blank or fill out the PUB section and set the status=1 (unpublished). Please see dbSNP's online submission instructions.

When submitting, what do I put down in the publication section if the paper is not yet published?

You can either leave this section out, or if you anticipate publishing later, define a publication section and set the status field as 1 (unpublished).

```
STATUS: Status field.
1=unpublished, 2=submitted, 3=in press, 4=published
```

The SAMPLESIZE field

What is the difference between the SAMPLESIZE field in SNPASSAY and the SAMPLESIZE I need to fill out for each SNP?

For SNPASSAY, SAMPLESIZE is the number of chromosomes you assayed for the entire batch. If you want to submit data for over 10,000 SNPs, we suggest that you create separate batches for each chromosome.

The SAMPLESIZE for each SNP is the number of chromosomes you sequenced to detect the variation. If you have 100 individual diploid samples, give 200 as the sample size.

In most cases, the two SAMPLESIZE(s) are the same, but occasionally SAMPLESIZE may be smaller in the SNP section than in the SNPASSAY when an experiment fails for some samples. (**04/14/08**)

In the submission form, what do I put for "SAMPLESIZE" If I sampled 100 people, and 27 of them have a different allele according to the NCBI consensus sequence?

Provide the number of chromosomes you sequenced to detect the variation. If you have 100 individual diploid samples, give 200 as the sample site. The Submission templates available on the dbSNP FTP size now include instructions for each field. You can also go to the dbSNP submission Quick Start for more information about filling out your submission form (**5/22/07**)

I want to submit SNP genotype data for four different populations, but don't understand the concept of a batch and what I use as the sample size for a batch.

We find it useful to use a single population per batch. The sample size is the number of chromosomes in the population for each batch. If you want to submit data for over 10,000 SNPs, we suggest that you create separate batches for each chromosome.(**10/2/06**)

Will the SAMPLESIZE field equal 416 if I assayed separate samples of diploid germline genomic DNA from a total of 208 individuals?

416 is correct, if the cells are diploid germline and not sperm or egg cells which are haploids. **(5/4/05)**

5'_ASSAY and 3'_ASSAY fields

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**)

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**)

What is the distinction between the 5'_Flank/3'_Flank and the 5'_ Assay /3'_Assay fields?

The ASSAY sequence is detected or sequenced by the experiment, and the FLANK is additional 3' or 5' reference sequence provided if the ASSAY is less than 100 bp. The additional FLANK is required by BLAST to aid in mapping the SNP to the genome. **(2/11/05)**

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**)

When filling out my submission form, which strand should I use for the 3'_ASSAY? Do I use the same strand as the 5_ASSAY, or the reverse compliment to the 5'_ASSAY sequence?

A description of strand orientation is located in dbSNP’s submission submission documentation. Unfortunately, however, the concept of strand orientation as used in dbSNP submissions is not the same as the concept of strand orientation that we all understand as molecular biologists.

Orientation tags, such as SS_STRAND_FWD, that are used in submitting SNP data, refer to the allele's direction relative to the flanking sequence's direction. It has nothing to do with the orientation of the gene on the contig. Therefore, in a dbSNP submission, you would use SS_STRAND_FWD if your allele falls on the strand that contains the forward orientation of your flanking sequence, and you would use SS_STRAND_REV if your allele falls on the strand that contains the reverse orientation of your flanking sequence.

You always report the allele as you observe it (your actual results). The strand_value, however, is determined by the primers you use to sequence. So if your primers attach to the strand that contains the forward orientation of your flanking sequence, use SS_STRAND_FWD. If your primers attach to the strand that contains the reverse compliment of your flanking sequence, you would use SS_STRAND_REV. (**3/21/07**)

5′_FLANK and 3′_FLANK fields

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**)

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**)

What is the distinction between the 5'_Flank/3'_Flank and the 5'_ Assay /3'_Assay fields?

The ASSAY sequence is detected or sequenced by the experiment, and the FLANK is additional 3' or 5' reference sequence provided if the ASSAY is less than 100 bp. The additional FLANK is required by BLAST to aid in mapping the SNP to the genome. **(2/11/05)**

Can I include more than100 total bases of flanking sequence in a submission?

1. The sum of the 5' and 3' sequences must be at least 100 bp.
2. The 5' and 3' sequences must be at least 25 bp each.
3. The total number of bases in the sum of: 5' + 3' + observed must equal less than 250 bp.

(**08/20/07**)

Do I include the assay sequence when I fill out the 5'_FLANK field in the submission form?

dbSNP maps submitted SNPs using the sequence of 5'_flank (+5'_assay) + observed + (3'_assay) + 3'_flank sequence. Inclusion of the assay sequences is optional, but if included, the assay sequences cannot overlap the flanking sequences. (**6/8/07**)

If we have a SNP that is less than 25 bases from the end of our sequence, should we submit it, even though the instructions say that the minimum flanking sequence either 5' or 3' must be at least 25 bases?

You can submit it if you want; we will load it, but I'm not sure it will map to the correct position. (**7/25/07**)

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**)