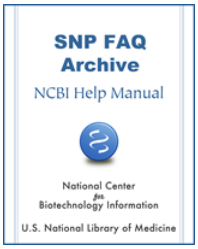
U.S. National Library of Medicine
National Center for Biotechnology Information

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



Sequence Formatting in dbSNP Reports

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

Sequence Orientation

Based on the Illumina note, am I correct in thinking that the same strand could be designated both "top" and "bottom" depending on which SNP was being examined?

You are correct in thinking that a long strand of DNA with many SNPs on it could have a different "top" or "bottom" designation at each SNP position based on Illumina's top/bottom rule.

Think of Illumina's top/bottom strand designation as a strand rule in the "LOCAL" (to a SNP) sense. It is useful when comparing genotype data for each SNP, especially those SNPs with no (A/T, C/G) alleles.

For an A/G SNP, if probe "A" produced a C/T result, and probe "B" produced an A/G result, then probe "A" *is a bottom strand probe* and probe "B" is a top strand probe. Notice that I said "*probe 'A 'is a bottom strand probe*" instead of "probe 'A' is *ON* the bottom strand".

To me, "ON" implies that a "top/bottom" designation applies to an arbitrarily long sequence, when in fact the "top/bottom" designation is just a way to compare each SNP genotype result. Again, the key phrase here is "LOCAL to a SNP".

Please use "orientation" flags to identify a sequence's relative orientation to a contig, mRNA or genome. (**08/08/08**)

How do I determine the top/bottom strand in the Illumina SNP array data?

First of all, be sure to read the ILLUMINA guide to their method for determining strand.

There are two ways you can get the top/bottom designation:

1. You can compute the top/bottom designation yourself using the data in the /organisms/human_9606/GWAS_arrays/ directory on the dbSNP FTP site.
2. You can look at dbSNP's top/bottom assignment, which you can access if you download the SubSNP.bcp file located in the /database/organism_data/ directory for human. The field that includes the top/bottom data is called SubSNP.top_or_bot_strand. You can access the table DDL for SubSNP in the /database/organism_schema directory.
   The downside of this approach is that you need to download the entire SubSNP table, which includes 50million+ submitted SNPs.

(**09/22/08**)

Will refSNP flanks change orientation between builds?

A refSNP's flanking sequence will never change orientation, but a refSNP's orientation with respect to the genome may change between builds if the genome assembly itself has significant changes that occur between

builds. This was the case in the earlier human genome builds, but the human genome build is more stable now, so orientation changes such as this will occur less often.

There are several different orientation types which exist in dbSNP:

- The orientation of a submitted SNP(ss) flank with respect to the RefSNP cluster (rs) flank.
- The orientation of the rs flank with respect to the contig sequence.
- The orientation of the contig sequence with respect to the genome.
  Please note that all placed human contigs are in the same orientation as genome.
- The orientation of a refSNP with respect to the genome.
  Since all placed contigs have the same orientation as genome, this orientation is the same as rs orientation to contig in human.
- The orientation of an mRNA with respect to the contig.
  This might not be related to our discussion here, but I mention it since it might come up in another context.

We make sure that a refSNP's flanking sequence orientation never changes: If a new ss is added to a refSNP cluster, and if that new ss has the longest flanking sequence (and therefore becomes the exemplar of the cluster), but has reverse orientation with respect to the existing rs, we reverse its flanking sequence when it becomes the new rs flank. (**11/20/07**)

rs135551 is marked reverse ("rev") in dbSNP, but is actually in forward orientation in the genome (chromosome 22). This orientation discrepancy is proving difficult.

rs135551 is a refSNP Cluster with 12 submitted SNP(ss) numbers. An ss number gets assigned to a refSNP (rs) cluster based on flanking alignment similarity. An ss number can be either in forward or reverse orientation with respect to its rs cluster. The "rev" in the submission section of the refSNP report shows the strand orientation of the member ss numbers in the cluster with respect to the refSNP.

If you look in the FASTA section of the report, you will see the flanking sequence for rs135551, with the sequence closest to the variation reading as follows:

TTAGACTCAG Y GAGGACAGTC

The above flanking sequence aligns to chromosome 22 in the reverse orientation on both the NCBI and the Celera assemblies.

I'm guessing that in your alignment to chromosome 22, you used an ss number within the cluster that had reverse orientation with respect to rs135551, and hence got your forward orientation result. (**10/17/07**)

How is "orientation" determined in the "Submitter Records" section of the refSNP report? Is it the submitted SNP orientation with respect to the plus and minus strands?

The orientation is the orientation of the submitted SNP with respect to the refSNP (either plus or minus strand of the refSNP). Sequences are always shown as 5'->3'.(**10/31/07**)

I am interested in retrieving flanking sequences in the forward orientation for a list of b126 SNPs. How do I do this?

A refSNP (rs) flanking sequence is simply the flanking sequence of the longest submitted SNP (ss) in the refSNP cluster. The ss with the longest flanking sequence is called the "refSNP exemplar". If a refSNP cluster gets a new ss member added after build 126 and this new ss has flanking sequence that is longer than the flanking sequences of the existing ss in the cluster, then the new ss becomes the refSNP cluster's exemplar, its flanking sequence is adjusted for orientation, and it will be used as the rs cluster's flanking sequence in the next build. Since the new ss will, in most cases, align at the same position as the rs, the flanking sequence

difference should be small. I am therefore curious why you want would need the rs flanking sequence for build 126. Have you noticed a significant difference (other than length) between rs flanks in different builds?

In general, dbSNP does not keep old build data due to data size issues and the complexity of tracking assembly changes between builds. However, if you have a local copy of dbSNP, you can access the rs flanking sequence for a particular build since dbSNP keeps the flanking sequences of all submitted SNPs. If you do not have a local copy of dbSNP that you can query, give us a list of the rs numbers in question, we can pull the data for you.(**11/20/07**)

Define the term "orientation" as used in dbSNP.

Submissions to our database have arbitrary orientation relative to each other. If multiple submissions refer to the same SNP, they may cluster together in reverse orientation, so we also track the orientation of each submission relative to the exemplar ss. Please bear in mind that submitters to dbSNP are only required to provide some flanking sequence around the SNP for context. The SNPdev team does the positioning using BLAST and the resulting alignments. **(3/13/05)**

How do I determine the orientation of dbSNP's allele frequency data?

On the refSNP page, allele frequency is always reported in the same orientation as the flanking sequence in the refSNP page FASTA section. When frequency data is submitted, we ask the submitter to specify strand information using tags like: SS_STRAND_FWD, SS_STRAND_REV, RS_STRAND_FWD and RS_STRAND_REV. If no strand tags are submitted, we assume the strand is in the same orientation as the submitted SNP or the refSNP. When computing refSNP allele frequency, we reverse the alleles when necessary. Sometimes frequency data is submitted for the wrong strand. If the alleles are A/T or C/G, we have no way of knowing that they have been submitted improperly. Please contact snp-admin@ncbi.nlm.nih.gov if you find any errors in frequency data.

The actual base change for rs3737085 is C>G, but the flanks reported in the "Submitter Records" section of the refSNP Cluster Report shows two other nucleotides in red, with no specific refSNP numbers assigned to them.

The red "c" and "t" in the "5' Near Seq 30 bp" and "3' Near Seq 30 bp" columns in the submitter records section indicate the bases used to determine the TOP/BOTTOM strand code as developed by ILLUMINA. The TOP/BOTTOM strand code is useful when determining the strand of genotype results. If you are interested, you can see a detailed description of the TOP/BOTTOM strand code online. If you are concerned about neighboring SNPs for rs3737085, and would like to see them, go to the Integrated Maps section of the refSNP page, and click on the word "view" located in the "Neighbor SNP" column. In this case, you can see that there is a SNP 12 bp away from rs3737085. (**7/20/07**)

I think "alleles" and "db SNP allele" may be switched in rs28944222, where dbSNP shows A/G; S, P; and in rs28944221 where dbSNP shows T/C; N, and D.

Both of these rs numbers mapped to the reverse strand of the contig, while the mRNA mapped to the forward strand:

==========> Contig [Forward]

----------> mRNA [Forward]

<------ SNP [Reverse]

You must therefore use the complementing nucleotides of the SNP alleles in order to get the correct codon, which will in turn, code for the correct amino acid:

T/C is the complement of A/G and codes for S, P

A/G is the complement of T/C and codes for N, D (**1/5/06**)

IUPAC Code

Are the M, Y, R, W, K and S codes you use in the FASTA sequence at the SNP position designed by American Association of Biochemistry?

Yes. We use IUPAC codes in FASTA sequences at the SNP position. You can find the IUPAC code in many websites. Here is an example of one such listing.(**07/22/08**)

Does dbSNP have a file that contains SNPs in IUPAC code for the entire human genome?

Sorry, there isn't a file with the variations encoded in IUPAC code for the entire human genome. **(1/10/05)**

What does N/N represent in refSNP sequence?

N/N is the IUPAC code used to indicate that the actual base can't be determined by a genotyping assay. **(11/17/05)**

I have come across nucleotide representations that I don't understand. What does "R" mean?

"R" is part of the IUPAC code for nucleotide variations which represents "A" or "G". You can find all of the IUPAC nucleotide codes online. **(3/9/05)**

Lowercase (Small) Sequence Lettering

We have found lowercase (small) letters in SNP sequence reports downloaded from your FTP site. What do these lowercase letters mean?

Lowercase (small) lettering is used for sequences identified by RepeatMasker as repetitive elements or as low complexity. You can find this information by accessing the refSNP page FASTA section and clicking on Legend.

In SNP flank sequence, I find that some bases are capital letters, while others are in lower-case lettering. What is the difference between the two?

Sequence in lower case has been identified by RepeatMasker as low-complexity or repetitive elements. You can find this description by clicking on the "Legend" link located above the sequence, which will take you to a sequence descriptions page. **(9/23/05)**

dbSNP 0-based (zero based) vs. 1-based Coordinate Representation

The XML dump for build 126 has a -1 coordinate error that has propagated to all records. Is this change intentional?

In order to meet NCBI guidelines, dbSNP changed the sequence coordinate storage and representation in the XML, ASN.1, .bcp, and the Genotype/ genotype_by_gene files from 1-based to 0-based starting with dbSNP build 125.

ASN.1_flat files, Chromosome Reports, and the web page reports remain 1- based. (**04/18/08**)

Flanking Sequence

Discrepancies in Flanking Sequence

The actual base change for rs3737085 is C>G, but the flanks reported in the "Submitter Records" section of the refSNP Cluster Report shows two other nucleotides in red, with no specific refSNP numbers assigned to them.

The red "c" and "t" in the "5' Near Seq 30 bp" and "3' Near Seq 30 bp" columns in the submitter records section indicate the bases used to determine the TOP/BOTTOM strand code as developed by ILLUMINA. The TOP/BOTTOM strand code is useful when determining the strand of genotype results. If you are interested, you can see a detailed description of the TOP/BOTTOM strand code online. If you are concerned about neighboring SNPs for rs3737085, and would like to see them, go to the Integrated Maps section of the refSNP page, and click on the word "view" located in the "Neighbor SNP" column. In this case, you can see that there is a SNP 12 bp away from rs3737085. (**7/20/07**)

Many bovine SNPs in dbSNP seem to have the first 10 bases of both flanks repeated again later in the flank. rs41567118 is just one example of many.

rs41567118 is a refSNP cluster, and as of this date, the cluster contains a single submitted SNP(ss61470123http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?searchType=adhoc_search&type=rs&rs=rs41567118). The original submission file for ss61470123 shows that the SNP was submitted using the 5'_ASSAY sequence as the last 10 bp of the 5'_FLANK. This submission error resulted from a misunderstanding of the submission format.

dbSNP maps subnmitted 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, they cannot overlap with the flanking sequences. In your example, the flanking sequences for ss61470123 overlap the assay sequences. We will notify the user to correct the problem and resubmit the variation. (**6/8/07**)

Ambiguity Symbols in Flanking Sequence

I noticed a refSNP (rs) that had several ambiguity symbols in the flanking sequence. I thought only SNPs themselves are described with ambiguity symbols.

A refSNP (rs) flank is simply the flanking sequence of the cluster member submitted SNP (ss) that has the longest flank. In this case, the submitted SNP from SNP500CANCER has the longest flank and they happened to use ambiguity symbols in their flanking sequences to denote neighboring SNPs.

You can look at these neighboring SNPs by scrolling down to the Integrated Maps section of the refSNP page, and then scrolling to the right to find the "neighbor SNP" column, and then clicking on the text link "view".

In the future, we plan to provide the refSNP flanking sequence with ambiguity symbols for neighboring SNPs. When this change is implemented, it will be broadcast in the dbsnp-announce email. (**5/8/07**)

Flanking Sequence Location

Is the sequence contained in each dbSNP record the parent sequence of the SNP, or does the sequence span a region just around the polymorphic site?

The displayed sequence spans a region around the polymorphic site

Neighbor Variations in Flank

The actual base change for rs3737085 is C>G, but the flanks reported in the "Submitter Records" section of the refSNP Cluster Report shows two other nucleotides in red, with no specific refSNP numbers assigned to them.

The red "c" and "t" in the "5' Near Seq 30 bp" and "3' Near Seq 30 bp" columns in the submitter records section indicate the bases used to determine the TOP/BOTTOM strand code as developed by ILLUMINA. The TOP/BOTTOM strand code is useful when determining the strand of genotype results. If you are interested, you can see a detailed description of the TOP/BOTTOM strand code online. If you are concerned about neighboring SNPs for rs3737085, and would like to see them, go to the Integrated Maps section of the refSNP page, and click on the word "view" located in the "Neighbor SNP" column. In this case, you can see that there is a SNP 12 bp away from rs3737085. (**7/20/07**)