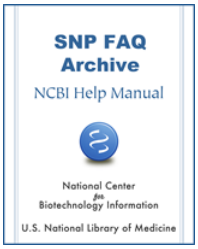
U.S. National Library of Medicine
National Center for Biotechnology Information

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




# Finding Information in a dbSNP Data Report

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

## Entrez SNP Graphic Summary of Search

**I've used Entrez SNP to search for SNPs on a gene, but the results (graphic summary) page shows no SNP positions. How do I find the correct position of the SNPs on the gene?**

Click on the "GeneView" (dark pink icon) located on the Entrez SNP graphic summary page below each refSNP (rs) number's sequence information. (**06/23/08**)

## refSNP (rs) Cluster Reports

### The Build Created/Updated Field

**dbSNP documentation indicates that the physical location for rs1256349 is from build 121, yet a phrase called Last updated in the build field of the rs1256349 report shows that this rs number was last updated in build 92.**

Last updated indicates the last build where the physical location of this SNP required an update. The value for Last updated would be changed if the physical location of the SNP was different between builds. In this case (rs1256349), the location of the SNP was stable between builds, and therefore the value for Last updated has not changed since build 92.

### The Alleles Field

**Why is rs67341913 reported as –/T in the cluster report allele section but annotated as**

**–/A in the RefSeq?**

This refSNP maps to the reference assembly on the forward strand, but the mRNA is on the reverse strand of the assembly, so the SNP allele is flipped (reverse complement) to '-/A' for reporting on mRNA.You can see this in the GeneView section of this rs number. (**06/08/09**)

**Does the order of the alleles listed in the alleles field of a refSNP report carry any biological meaning –that is does the "C" in the listing "C/A" mean that "C" is ancestral or wildtype or something?**

dbSNP does not distinguish between ancestral and wild type alleles in the way the variation alleles are listed in dbSNP. dbSNP simply lists the alleles in alphabetic order because:

- Many SNPs in dbSNP do not have ancestral allele information
- Some SNPs have different minor alleles between different populations
- dbSNP maps SNPs to all major assemblies (e.g. dbSNP maps to NCBI reference, Celera and HuRef for human) and sometimes the different assemblies have different alleles at a particular SNP position.

**Therefore, the order of the alleles does NOT carry any biological meaning**.

If the SNP has frequency information, you can check the frequency data over each specific population to infer whether the allele is wild type or mutant in a specific population. For information to help you determine ancestral alleles, please see the ancestral allele section of the SNP FAQ Archive. (**09/23/08**)

**There is a link to "VarView" in the Allele column of the cluster report for a number of refSNPs I'm looking at. What exactly is "VarView"?**

"VarView" (short for "Variation Viewer") icons or links result in a gene-specific display named for the variation view of the gene in question: "[Gene Symbol]" + "Variation Viewer" (e.g. MECP2 Variation Viewer).

VarView is an improved alternative to dbSNP's GeneView in that it contains more intuitive packaging of the data found in dbSNP records (e.g. HGVS names, numbers of observations, clinical associations, links to OMIM and Locus-specific databases (LSDB), citations, etc.).

Currently, a VarView link or icon appears in a variation record only if the variant was submitted with clinical association(s), and if the gene in question has a RefSeqGene record.(**09/22/08**)

**Why do web queries of rs939820, rs10205833, and rs7597158 return sequences that do not match the exemplar sequences for these SNPs found using a database query?**

When you refer to the sequences as "not matching", I assume that you are referring the fact that they don't match at the point of variation, since in rs939820, for example, the two flanking sequences are the same with the exception of the point of variation.

On the RefSNP (rs) page, we show the rs FASTA using the IUPAC code for variations, while on the submitted SNP (ss) page, we show the ss fasta using the submitted observed sequence. In most cases, all member ss of an rs cluster have the same allelic states in the same orientation, so the rs variation matches the ss exemplar variation. There are cases, however, where the rs variation does not match the ss exemplar variation. For example, if an ss exemplar has an A/G variation, and another ss from the cluster in the same orientation has an A/T variation, then the rs allele list will read A/G/T since it includes all member ss alleles. If you viewed the rs allele list converted into IUPAC code (remember the refSNP page shows the flanking sequence in IUPAC), it would show a D representing A or G or T.

Most of the submitted SNPs have the same variations in the rs clusters you mentioned, but one or two of the ss in each cluster have an extra allele. All of the ss in the rs939820 cluster have an A/G variation, with the exception of one ss that has an -/A/G variation. All the ss in the rs10205833 cluster have a C/G variation, with the exception of one ss that has a C/G/T variation. Most of the ss in the rs7597158 cluster have an A/G variation, while the ss exemplar has a -/G variation. In these cases, the refSNP page variation list includes all allelic states: for rs939820, instead of an R, it is N; For rs10205833, instead of an S, it is B that represents for C or G or T; for rs7597158, since the ss exemplar has a -/G variation, while most of the other ss in this cluster have an A/G variation, the refSNP FASTA shows an N at the variation point.

I will update dbSNP's variation representation for "mixed variation" clusters to include all the allele lists from all the submitted SNPs.(**8/18/06**)

## Frequency

**How many times was the variation represented by rs8137714 observed? What population and how many chromosomes did you screen?**

Currently, rs8137714 does not have genotype or allele frequency information in dbSNP. To see how the submitted SNPs in this refSNP cluster were mined, do the following:

1. Go to the refSNP report for rs8137714 and look at the "Submitter Records" section

2. Click on either of the NCBI Assay ID numbers located on the left. This will take you to the submitted SNP Detail report for the selected SNP.
3. Scroll down to the Assay section of the Detail Report, and click on the assay "Method" listed there. This will take you to a description of the method used for finding the submitted SNP.

By following the above steps, you can see that both submitted SNPs in this cluster were computationally mined.

If you look at the Validation Summary" section located at the bottom of the refSNP report, you will see that this SNP is a "Double Hit" SNP. In other words, "all alleles have been observed in at least two chromosomes apiece." You can find more information on Double hit SNPs in the dbSNP FAQ Archive.

Some submitters will submit frequency information for existing refSNP numbers, so it is possible that dbSNP will get frequency information for this refSNP (rs) number in the future. (**11/26/07**)

**The reported frequency of a minor variant is GCGATTGGCC{G/A}GGACCACGAC 0.102 plus/minus 0.012. Can I assume that the minor variant is A?**

When a SNP is submitted as (G/A), we do not assume that the "G" allele is the major (wild-type) variant, and the "A" allele is the minor variant. Because the minor allele in one population can be the major allele in an other population, we do not assume allele order implies allele frequency information, and do not interpret the data at all — to bdSNP, both the "G" variant and the "A" variant have been observed at the SNP location, and that is all.

At the same time, however, we have been, and will continue to encourage dbSNP users to submit individual genotype data or genotype/allele frequency data for both new and existing SNPs. Once we have allele frequency data, we can then state the minor allele and its frequency. (**5/1/06**)

## Strand Orientation

**How do I determine orientation of a SNP allele (rs9934438) in dbSNP, and then find the corresponding strand and position information in the UCSC genome browser?**

To find orientation information about a SNP, go to the "Integrated Maps" section of the SNP's cluster report. In the case of rs9934438, you can see that rs9934438 hits on the plus strand of NT_010393.15 in the reference assembly. rs9934438 has allele "A/G" (see the "allele" section at the top of the page) and the contig allele of NT_010393.15 is "G" at the SNP position.

To find the link to UCSC's genome browser for this refSNP, click on "Links" (located just to the right of the "Alleles" section at the top of the refSNP page) and choose "UC Santa Cruz" to go to the UCSC browser at the location of this particular SNP. As of this date, B129 is the most recent.

(**09/12/08**)

**dbSNP reports rs10512248 alleles as A/C, but Nature Genetics 40(5):pp.575 table 1 reports rs10512248 as G/T, where "alleles all refer to the positive strand."**

I think the "positive strand" mentioned in this paper refers to the NCBI reference assembly strand.

rs10512248 maps to the NCBI reference sequence on the minus strand (see the integrated maps section of the cluster report). That is why dbSNP shows the rs allele as reversed to what the paper reports. In other words, SNP alleles can be described as either in "RefSNP orientation" or in "reference genome strand orientation". If you take the orientation of both the refSNP and the genome into consideration, then the data from dbSNP and the data in the paper are consistent.

Here are a couple of reasons dbSNP does not report alleles in genome orientation:

- dbSNP maps refSNPs to several alternative assemblies (Celera, HuRef) as well as to the reference assembly, and sometimes these different assemblies are in different orientations at the SNP position.
- Some SNPs do not map to any genome positions or they may have multiple positions, making it impossible to report the refSNP allele in genome orientation.

I would also like to point out that a refSNP never changes orientation between dbSNP builds. And if we assume genome sequences do not change strand orientation (which is, for the most part, true), then if a refSNP maps to the minus strand of build 35, then it will also map to minus strand in build 36, and to the minus strand in other future builds.(**08/06/08**)

**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**)

**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**)

**In the refSNP report for rs1805794, does the allele notation "C/G" mean that C is on the forward strand of the chromosome in the reference assembly (and the variation is C/G)?**

Alleles as shown in the RefSNP report are in the order dictated by the ascii character set. C/G, therefore, does not imply that the C is on the forward strand of chromosome in reference assembly. The "Integrated Mapping" section of the report shows the contig allele and orientation. In this case, the contig allele is "C" on the reference assembly, and it is in the same orientation as the SNP flanking sequence. (**5/4/06**)

**How do I find the set of alleles that you used to instantiate a SNP allele sequence?**

We do not choose the alleles for a SNP. We include all alleles reported by submitters for a refSNP cluster. If a submitted SNP is on the reverse strand relative to the refSNP sequence, we reverse the alleles. For example, in the refSNP(rs) cluster rs268, rs268and ss268 have the allele A/G, while ss48420135 shows the C/T allele on the reverse strand. So the reported allele set for refSNP cluster rs268 is reported as A/G.

(**6/21/06**)

## Variation Position Notation

**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**)

**In one place, the report for rs6993291 shows the variant allele as M, and then elsewhere in the report it says the variant allele is A/C. Which is it? What does M stand for?**

In some cases the report may display the IUPAC ambiguity code equivalent of the variant allele rather than displaying the variant allele itself (e.g. M = A/C). (**3/7/06**)

**The variation position notation for rs1922237 is: "12382526^12382527". What does this notation mean?**

In this case, the contig has a deletion at the SNP position and actually has a new "allele" (a deletion) that was not reported by the submitters to dbSNP.

When a contig has a deletion at a SNP position, it does not necessarily mean that the SNP is a DIP (**D**eletion **I**nsertion **P**olymorphisim), since dbSNP does not always have an exhaustive list of variation alleles for each SNP.

The SNP to contig positions in SNP builds so far have three loc_type(s), each of which has a specific notation:

- loc_type 1, where the contig has multiple bases. The range for loc_type 1 is: "asn-from < asn-to", which is written as "asn-from..asn-to".
- loc_type 2, where the contig has exactly one base. The range for loc_type 2 is: exact, where "asn-from = asn-to", and is written as "asn-from"
- loc_type 3, where the contig has a deletion. The range for loc_type 3 is: between, where "asn-to = asn-from+1", and is written as "asn-from^asn-to". This loc_type does not depend on the SNP variation class, whether it is a true SNP or a DIP. **(9/19/05)**

**The variation position notation for rs10600037 is: "DIP(+) seq 12489885..12489886". What does this notation mean?**

If there are multiple bases at a SNP position on the contig, then ".." is used to show the location of the variation. In this example, rs10600037 has the deletion insertion polymorphisim (DIP)" —/TG, where the two bases "TG" align at the SNP variation position: 12489885..12489886 on the contig.

The SNP to contig positions in SNP builds so far have three loc_type(s), each of which has a specific notation:

- loc_type 1, where the contig has multiple bases. The range for loc_type 1 is: “asn-from < asn-to”, which is written as "asn-from..asn-to".
- loc_type 2, where the contig has exactly one base. The range for loc_type 2 is: exact, where “asn-from = asn-to”, and is written as "asn-from"
- loc_type 3, where the contig has a deletion. The range for loc_type 3 is: between, where “asn-to = asn-from+1”, and is written as "asn-from^asn-to”. This loc_type does not depend on the SNP variation class, whether it is a true SNP or a DIP. **(9/19/05)**

**The variation position notation for rs10532925 is: “DIP(+) seq 86843313”, but this notation does not have the “^”or “..” I’ve seen in the variation position notation for other refSNPs. What does this notation mean?**

When there is one base at the SNP position on the contig, then just the base position is listed. rs10532925 has the deletion insertion polymorphisim (DIP)”—/AAA, but the contig allele is "A". So this is a case of the contig showing a new allele for the SNP.

The SNP to contig positions in SNP builds so far have three loc_type(s), each of which has a specific notation:

- loc_type 1, where the contig has multiple bases. The range for loc_type 1 is: “asn-from < asn-to”, which is written as "asn-from..asn-to".
- loc_type 2, where the contig has exactly one base. The range for loc_type 2 is: exact, where “asn-from = asn-to”, and is written as "asn-from"
- loc_type 3, where the contig has a deletion. The range for loc_type 3 is: between, where “asn-to = asn-from+1”, and is written as "asn-from^asn-to”. This loc_type does not depend on the SNP variation class, whether it is a true SNP or a DIP. **(9/19/05)**

**The variation position notation for rs4148752 is: “DIP(-) seq 12376835^12376836”. What does this notation mean?**

rs4148752 has the deletion insertion polymorphisim: “—/AAAG”, where the contig has the deletion [thus, “DIP(-)”]

The SNP to contig positions in SNP builds so far have three loc_type(s), each of which has a specific notation:

- loc_type 1, where the contig has multiple bases. The range for loc_type 1 is: “asn-from < asn-to”, which is written as "asn-from..asn-to".
- loc_type 2, where the contig has exactly one base. The range for loc_type 2 is: exact, where “asn-from = asn-to”, and is written as "asn-from"
- loc_type 3, where the contig has a deletion. The range for loc_type 3 is: between, where “asn-to = asn-from+1”, and is written as "asn-from^asn-to”. This loc_type does not depend on the SNP variation class, whether it is a true SNP or a DIP. **(9/19/05)**

## The “HGVS Names” Field

**The HGVS name for rs6729072 is NM_012233.1:c.1067-44G>T; I expected from this name that G would be the major allele and T the minor allele, but the cluster report showed the opposite.**

According to HGVS nomenclature, the allele before the ">" sign is simply the base located on the reference sequence at the position specified in the HGVS name — it does not have to be the major allele.

The major/minor allele is a concept of population genetics, and is specific to a certain context, such as a particular population.

(**06/03/09**)

## The Submitter Records Section

**dbSNP's "Validation status description" states that one mode of validation is when "All alleles have been observed in at least two chromosomes apiece." Do you mean chromosomes from different populations?**

There is a detailed discussion of the double hit validation in the "Build Process" section of the archive. (**10/25/07**)

**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**)

**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**)

**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**)

**What does it mean when a SNP (rs627099) has a symbol indicating a "two hits" validation status in the submitter records section of the refSNP report?**

If you click on the column header "Validation Status" in the refSNP page, you will get a brief description for each validation status symbol. The "two hit" symbol means that all alleles have been observed in at least two chromosomes apiece.(**10/31/06**)

**What is the definition for the "Orientation/Strand" field in SNP cluster reports and what are the definitions for the field values "F/T","F/B","R/T","R/B"?**

You can find definitions for these terms (described below) by clicking on the "Orientation/Strand" column header in the Submitter Records section of the refSNP cluster report:

Orientation: F = forward, R = reverse.

When in the context of clustering submitted SNPs, "F" means in the same orientation as the refSNP sequences.

Strand: T = top. B = Bottom. This notation was proposed by ILLUMINA and is detailed in an ILLUMINA document located on the dbSNP FTP site. **(9/15/05)**

**How do I determine who submitted rs11544612?**

Go to the rs11544612 refSNP page and scroll down to the "Submitter Records for this refSNP Cluster" section. Click on the text that reads: "ss16240609" (the only submitted SNP for this cluster) to go to the "Submitted SNP Detail" page for this ss number. Once you are on the detail page, click on the text for the submitter handle, which in this case is "CGAP-GAI", to go to the "SNP Submitter Contact Detail" page, which contains all full name of the submitting individual or institution along with their contact information. (**9/7/06**)

## The FASTA Sequence Section

**Do you have a key for the refSNP cluster report FASTA section that would help me decipher the sequence colors and the information above the sequence?**

Next to the FASTA section header in the blue section bar, there is the word "Legend" in parenthesis. Click on it to see a key that will define the terms and colors. **(03/26/08)**

**Some of the sequence in the FASTA section of rs6265's cluster report is green and some is black. What do these colors mean?**

If you click on the word "Legend" Located in the FASTA section divider (blue bar), you will get a drop-down menu. Scroll to the bottom of this menu, and you will find the following definitions:

- The Green color is used for assay sequence (observed by the submitter).
- The Black color is used for flank sequence (extracted from sequence databases).

(**02/15/08**)

## The GeneView Section

**There is a "has frequency" button In the GeneView section of the cluster report. Does this mean that there is either genotype and/or allele frequency data available?**

The button means that there is allele frequency data for the SNP. Either the frequency was submitted to dbSNP or we compute the frequency from submitted genotype data. (**06/18/09**)

**Why does the rs2522833 Geneview display show that an allele change of TCT to GCT when the allele section at the top of the cluster report shows A/C?**

**Please note** that the GeneView display shows the allele change in the mRNA field.

rs2522833 maps to the reference genomic assembly on the forward strand with alleles (A/C), but the mRNA is on the reverse strand of the assembly, so the SNP allele is flipped (reverse complement) to (T/G) for reporting on the mRNA.

(**06/23/09**)

**Where can I get a complete list of coding SNPs in a gene (ABCA4)? The GeneView page does not display all of them.**

There are quite a few SNPs on this gene (ABCA4, gene_id=24) that are from the LSDB (Locus Specific Database). These LSDB SNPs are considered "clinically associated", and the LSDB submitters prefer we do not show clinically associated rare variations as a default on the SNP GeneView page. You can include these clinical SNPs by clicking the "Include clinically associated" checkbox on the GenveView page, and then clicking the "Refresh" button located at the end of the same line where you found the "Include clinically associated" checkbox. If you looked at the "Refreshed" page which now includes clinically associated SNPs, you will find rs55732384, rs6657239 and rs58331765 are now present.(**10/15/08**)

**In the GeneView section of the cluster report for rs1137070, One of the "Group labels" is HuREF. What does HuRef mean?**

"Huref" is the Venter diploid genome. You can see the publication for this genome online. A quote from the "Genome Sequencing and Assembly" section of the paper defines HuRef in the following way:

"The assembly, herein referred to as HuRef, was derived of approximately 32 million sequence reads (Table S1) generated by a random shotgun sequencing approach using the open-source Celera Assembler." (**07/04/08**)

**It seems like all the disease related "clinically associated" links are missing from the "SNP linked to Gene via Contig Annotation" page for HFE.**

While on the "SNP linked to Gene via Contig Annotation" page for HFE, did you click on the "Include clinically associated" checkbox? It is located in a blue bar that separates the "mRNA alignment" section (first section) of the page from the "contig mRNA Transcript" section (second section) of the page.

We added this check box because some Locus Specific Database/Clinical SNP submitters did not want to see their rare variations mixed in with polymorphisms. (**06/02/08**)

**A rs300 query returns three contigs, and for each contig a different chromosome location. Which is the "right"one, and how do I find the position of the SNP on the corresponding gene?**

SNPs are not only mapped to the reference assembly, but also to alternative genome assemblies, hence the refSNP cluster report contains multiple contigs — one from each assembly.

Use the SNP contig position for the genome assembly that you are interested in. If you don't have a preference as to genome assembly, then I would suggest using the contig positions from the reference assembly.

The position of the SNP on the gene is shown in the "Integrated maps" section of GeneView section in the refSNP cluster report. The position reported will be assembly specific, so make sure you verify the assembly you are looking at by checking the "Contig Label" column.(**01/11/08**)

**Why is the sequence of gene IL18 (ID:3606) reversed — that is with the 5' end at the right and the 3' end at the left?**

The IL18 gene is on mRNA NM_001562.2, **which is on the reverse strand of contig** NT_033899.7. The SNP gene view page shows the contig from 5' to 3', and since the mRNA is on the reverse strand, it shown with the 3' end on the left and the 5' endon the right. (**12/28/07**)

**The GeneView section of the rs503351 reSNP report shows transcript XM_930578, but the GeneView page of XM_930578 doesn't show rs503351.**

rs503351 has been annotated on 11 genes; 10 of these genes have only interim gene names, such as "LOC#" , where "LOC" is an acronym for "locus ID", and "#" represents the gene ID number.

Go to the "GeneView" section of the refSNP report for rs503351, and scroll down until you find the entry "**GeneView via analysis of contig annotation:** LOC647135". Beneath this entry there is a viewing choice; choose "all". This will take you to the gene model which provides a list of all the refSNP IDs associated with

the gene, and their location on the gene. By scrolling down the page slightly, you'll find that rs503351 is in intron 1.

I can see in the refSNP report for rs50335 that the link between the gene name and the gene transcript is not clear when a refSNP is annotated on multiple genes. I will discuss this issue with the SNP development team to see if we can come up with some options that might clarify the presentation of the link between the gene name and its transcript in the future. (**9/5/06**)

**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**)

## The Integrated Maps Section

**How do I determine orientation of a SNP allele (rs9934438) in dbSNP, and then find the corresponding strand and position information in the UCSC genome browser?**

To find orientation information about a SNP, go to the "Integrated Maps" section of the SNP's cluster report. In the case of rs9934438, you can see that rs9934438 hits on the plus strand of NT_010393.15 in the reference assembly. rs9934438 has allele "A/G" (see the "allele" section at the top of the page) and the contig allele of NT_010393.15 is "G" at the SNP position.

To find the link to UCSC's genome browser for this refSNP, click on "Links" (located just to the right of the "Alleles" section at the top of the refSNP page) and choose "UC Santa Cruz" to go to the UCSC browser at the location of this particular SNP. As of this date, B129 is the most recent.

(**09/12/08**)

**The population diversity section of the rs11545130 cluster report shows an A-to-G change, but the integrated maps section shows a C nucleotide.**

You are correct that the cluster report shows the variation in two different orientations. The integrated maps section of the cluster report shows the variation in the reference sequence (NT_011520.11) orientation, while the population diversity section reports the frequency and allele data in refSNP orientation (the orientation of the cluster exemplar), which is why it was reported as A/G.

There are a number of reasons that we need to report RefSNP data based on RefSNP orientation defined by rs FASTA rather than based on reference sequence orientation:

- Sometimes a SNP maps to multiple reference sequences each of which have different orientations.
- Sometimes a SNP doesn't map to any genome position.
- Sometimes the NCBI reference sequence and the Celera or HuRef contigs are of different orientations.
- Between genome builds, contig sequences may change between different versions (although these changes becoming more rare as the contigs mature).

(**08/01/08**)

**Where can I find a reference that will show me the difference between an NT_ accession, a NG_ accession, and a NC_ accession?**

Table 1 in the "Database Content" subsection of the RefSeq section of the NCBI handbook contains a table defining "The RefSeq accession number format and molecule types". (**04/02/08**)

**A rs300 query returns three contigs, and for each contig a different chromosome location. Which is the "right"one, and how do I find the position of the SNP on the corresponding gene?**

SNPs are not only mapped to the reference assembly, but also to alternative genome assemblies, hence the refSNP cluster report contains multiple contigs — one from each assembly.

Use the SNP contig position for the genome assembly that you are interested in. If you don't have a preference as to genome assembly, then I would suggest using the contig positions from the reference assembly.

The position of the SNP on the gene is shown in the "Integrated maps" section of GeneView section in the refSNP cluster report. The position reported will be assembly specific, so make sure you verify the assembly you are looking at by checking the "Contig Label" column.(**01/11/08**)

**How do I identify proximal SNPs located around a primary SNP?**

You can find SNPs that are proximal to a given SNP by using the "Neighbor SNP" link located in the Integrated Maps section of the refSNP Cluster Report. For example, you can find the SNPs proximal to rs2515644 by scrolling down the cluster report until you came to the "Integrated Maps" section, which is located approximately half way down the page. Looking to the right, you'll see the "Neighbor SNP" column. Click on the blue "View" links located in the column to view neighboring SNPs. (**10/13/06**)

**Clicking on SNPs located in the "Graphic display" for the reference mRNA model, shows most of the SNPs to be introns, yet the graphic shows them all as coding. Why?**

The problem appears to be that the results of the two mapping pipelines for the SNPs didn't agree. One pipeline maps the SNPs to the genome (results are shown in the Reference Cluster Report), while the other pipeline maps the SNPs to the mRNA (results are shown in the mRNA graphic display). Although such a case is rare, it appears that some of the genomic SNPs (ie. rs2534719) map at the exon/intron boundary, and can therefore be classified as intron or exon depending on whether they map to the genome or the mRNA. **(5/26/05)**

**Why do maps associated with refSNP cluster reports, exclude regulatory sequences?**

SNP annotation is based on a SNP's co-location in the region of the genome that has been annotated with a gene. Currently, promoters and regulatory regions are not annotated on the genome, so we cannot provide SNP mapping to these regions. I think the problem lies in the difficulty of predicting promoter regions and validating them on a genome-wide scale. **(5/24/05)**

**On the refSNP page, where do I find information about amino acid changes resulting from the variation?**

When a SNP causes an amino acid change, you can see the changes in the Gene View section of the refSNP report (example). The Gene View section is located about half way down the refSNP report page. (**1/13/06**)

**I'm using the Sequenom iPLEX system for genotyping, and need 200bp of 5' and 3' flanking sequence for rs41296860. How do I find this much flanking sequence?**

Currently, dbSNP does not have data for long flanks as you request, so you will have to use the following steps to find the

rs41296860 flanking sequence you need:

1. Go to the Integrated Maps section of the refSNP cluster report for rs41296860.
2. Click on NT_021877.18. This will take you to an Entrez Nucleotide report showing rs41296860 and its flanking sequence on the genomic contig
3. Since the variation's position spans the contig (NT_021877.18) between 467055 and 467056, all you have to do is subtract or add the number of bases you want from the variation positions, and type the

difference/sum in the range: "from" and "to" boxes (e.g. 467055 – 100 =466955 and 467056 + 100 = 467156 for a flank of 100bp on either side of the variation).
4. Click the "Refresh" button to see the fasta sequence in the desired length.

(**04/04/08**)

## The NCBI Resource Link Section

**Does each dbSNP entry have a link that leads to the parent sequence in GenBank?**

Submitters often supply the source sequence, but it is not required. For example, let's look at rs12345 using the Reference SNP (refSNP) Cluster Report (this is where we cluster redundant submissions). If you scroll down to NCBI Resource Links, you will see a list of GenBank accessions under Submitter-Referenced Accessions:. In this case, "Genbank: AA854219 AL008721 BU602737 Hs.169286" is the union of all submitted GenBank accessions for all of the submitted SNPs in the cluster. Coordinates on the accession are not usually provided by the submitter. We BLAST alignments using the flanking sequence, and the resulting accession list is found just below, in dbSNP Blast Analysis.

## The Population Diversity Section

**Which tables and columns in the schema contain the data found in the Population Diversity section of the refSNP Cluster Report?**

The three values below, found in the refSNP Population Diversity section, are from tables in the SNP annotation area of the schema pdf:

Hardy-Weinberg Probability is located in the SNPHWProb table.
Average estimated heterozygosity is located in the SNP table.
Average Allele Frequency is located in the SNPAlleleFreq table.

Assay sample size (number of chromosomes), Population data sample size (number of chromosomes), Total number of populations with frequency data, and Total number of individuals with genotype data are all located in the refSNP page Population Diversity section, and are all computed on the fly.

**How can I download a tab-delimited file with Population Diversity information for hundreds of SNPs?**

There isn't a tab delimited report format containing the data that you requested. You'll have to upload your list of SNPs to the batch query service to get the XML or ASN report and parse out the data you want.

**How do I download the data located in the Population Diversity section of the refSNP report for all human SNPs? Would I get it on your FTP site?**

You can get the data from the following three database table dump files located in the dbSNP FTP Site:

SNP.bcp.gz (the relevant fields are: snp_id is the rs#, avg_heterozygosity)

SNPAlleleFreq.bcp.gz

SNPGtyFreq.bcp.gz

Also located in the ftp directory are the Allele.bcp.gz and the UniGty.bcp.gz tables, where you can find the meaning of each allele_id and/or unigty_id. The column descriptions for all the tables can be found in the dbSNP Data Dictionary. **(4/1/05)**

### Sample Ascertainment

**In the Population Diversity section of a refSNP page, what is the definition of "IG" as a "Source" in the Sample Ascertainment section?**

IG stands for "Individual Genotype". This means that dbSNP has the actual genotypes, not just the frequencies. (**7/19/07**)

**What is the definition of the population described in a SNP report?**

A "population" is submitted and defined by a submitter, which means that the submitter provides a population ID (an abbreviation) as well as a description of the population, therefore, “populations” are not curated by dbSNP.

Please note that genotype data also have populations, but the "population with frequency data" designation is only for populations that have allele frequency submissions, and these do not include genotype submissions.

You can see how major contributors to dbSNP define their populations by going to the “Search/View Population Detail” page and type in the submitter handles "Perelgen" and "CSHL-HapMap" in the search text box located in the grey search section and clicking the “Search” box directly below. **(3/9/05)**

**The report for rs1079610 shows an assay sample size of 65, whereas one of the ss reports for this cluster shows a 2N sample size of 1476. Why are the sample sizes different?**

There are two sample-size fields in dbSNP. One field is called the SNP Assay sample size; it reports the number of chromosomes in the sample used to initially ascertain or discover the variation. The other sample-size field is called SNPPOPUSE or population sample size; it reports the number of chromosomes used as the denominator in computing estimates of allele frequencies by dbSNP submitters. These two measures need not be the same.

In this case, the assay sample size for the refSNP is simply the addition of the sample sizes for each member ss. In this case: 48 (from YUSUKE) + 10 (from TSC) + 3 (from SC_JCM) + 2 (from SC_SNP) + 2 (from SSAHASNP) = 65. If multiple ss from the same submitter are in the same cluster, dbSNP counts the sample size only once for each submitter.

The population sample size of 1476 is the sample size reported with the population frequency data in ss4926339, reported by YUSUKE.

**The refSNP report for rs7735338, shows that there are 120 samples for the European population. Does this mean that the sample is 120 CEPH cell lines, or does it mean that there were 60 founders and two samples per founder?**

120 chromosomes from 60 diploid founders.(**5/11/06**)

**A refSNP report shows 90 individuals were sampled for genotype data. Is 90 the total number of individuals contributing chromosomes for the assay sample size and the population data sample size?**

90 is the total number of individuals contributing to individual genotype data. Please note that this number is independent of the assay sample size and the population data sample size since the person(s) who submitted the genotype data may not be the same person who submitted the original assay. In theory, it is possible that the same individuals were used in the discovery assay and for the genotype data, but dbSNP does not always have information with regard to this. **(3/9/05)**

## The “Source” Column

**The “Source” column in the Population Diversity section of the refSNP cluster report contains the abbreviations IG, AF, and GF. What do these abbreviations stand for?**

IG, AF and GF stand for Individual Genotype, Allele Frequency and Genotype Frequency, respectively. These three designations comprise the entire number of "population diversity" source types submitted to dbSNP. (**01/04/08**)

## Genotypes

**How many times was the variation represented by rs8137714 observed? What population and how many chromosomes did you screen?**

Currently, rs8137714 does not have genotype or allele frequency information in dbSNP. To see how the submitted SNPs in this refSNP cluster were mined, do the following:

1. Go to the refSNP report for rs8137714 and look at the “Submitter Records” section
2. Click on either of the NCBI Assay ID numbers located on the left. This will take you to the submitted SNP Detail report for the selected SNP.
3. Scroll down to the Assay section of the Detail Report, and click on the assay “Method” listed there. This will take you to a description of the method used for finding the submitted SNP.

By following the above steps, you can see that both submitted SNPs in this cluster were computationally mined.

If you look at the Validation Summary” section located at the bottom of the refSNP report, you will see that this SNP is a “Double Hit” SNP. In other words, “all alleles have been observed in at least two chromosomes apiece.” You can find more information on Double hit SNPs in the dbSNP FAQ Archive.

Some submitters will submit frequency information for existing refSNP numbers, so it is possible that dbSNP will get frequency information for this refSNP (rs) number in the future. (**11/26/07**)

**How is HWP (Hardy Weinberg Probability) found in the Population Diversity section of the refSNP report determined, and how does it relate to genotype data?**

A good explanation of HWP and how it is calculated can be found online. The FAQ Archive also contains information you should find useful. (**6/4/07**)

**Is the heterozygosity score (het score) for rs627099 derived from all four populations? Is there a Han Chinese heterozygosity score available?**

Yes, the het score is computed using all available population data. Look in the "Total Samples" (last) row of the Population Diversity section of the refSNP report.

Since the genotype is available for all HapMap samples, the "heterozygosity score" is just the frequency of the "heterozygous genotype". In your example of rs627099, the het score for (Han Chinese) HCB is 0.533, which is the freq for "C/G".(**10/31/06**)

**The refSNP report for rs2839858 doesn’t provide the frequency of each allele and the population diversity section of the report shows that all individuals have the A/A genotype. How can this be a SNP if there are no other observed alleles?**

Frequency and genotype data is not required in order to submit SNP assay data, although dbSNP does encourage its submitters to provide frequency and genotype information when submitting a new SNP.

Sometimes SNPs are discovered using a computational algorithm, and in such a case, there are no frequency and genotype data available. Other submitters, however, may submit the frequency/genotype data for such a SNP at a later date. In the case of rs2839858, you can see in the “Population Diversity” section of the report that the exemplar submitted SNP (ss4021194) genotype data was generated by HapMap. It is hard to say if this is truly a "Novariation" site. It might be a rare allele.

I encourage you to contact the submitter directly for more information. You can get the submitter’s contact information by clicking on a submitted SNP number (ss#). For example, if you click on ss48294011, you’ll go to the detail page for that SNP which contains the submitter information. Once on the detail page, click on the submitter’s handle, in this case "SNP500CANCER" to get the contact information.

You can also click on the handle/ submitted SNP ID located in the "Submitter Records" section of the refSNP report to get information on the method of discovery directly from the submitter. For example, if you click on "SNP500Cancer ID|DIO2-05" for ss48294011, you'll go to the SNP500CANCER page , which contains further details about this particular submitted SNP.(**9/14/06**)

**The refSNP report for rs7503991 shows an A/T variation, but the population diversity section shows the variation was 100% T in all populations. Why is this is listed as a SNP if only 1 allele is observed in all populations?**

This refSNP cluster was based on a single submission, which you can look at by clicking on the submitted SNP (ss) number located in the "Submitter records for this refSNP Cluster" section of the refSNP report, which will take you to the . Submitted SNP Detail Report.

The Submitted SNP Detail Report shows that this submission was based on a computer algorithm called SsahaSNP. If you would like to see the publication that describes the SNP discovery method, go to the Assay section of the Submitted SNP Detail Report and click on the method id number to see a description of the method.

The fact that the SNP was shown to be monomorphic in the four populations tested by HapMap could mean a few things. It could mean that there is no variation at this site since the computational method has its limitations. It could also mean that this is a rare SNP that is not present in the four populations tested. (**7/3/06**)

**The refSNP report for rs17362 shows the two genotypes for this refSNP listed simply as "A" and "C". Why do these genotypes look like this rather than the typical "C/C", "A/A", "A/C"?**

Since rs17362 maps to the Y chromosome, a genotype of "A" or "C" makes sense. All the genotype data submitted by Oefner has only a single letter. To see Oefner's submission in submission format, click on the submitted SNP (ss) number on the refSNP page to go to the submitted SNP detail report. Once you are there, go to the "Resource Links" section of the detail report (the located on the top right-hand side of the page), where you'll find the text "Submission Report" located at the bottom of the section. Click on the blue text "View" located next to the right of the "Submission Report" text.(**7/28/06**)

**I have encountered single letter genotypes for SNPs in XML files that are not located on chromosome Y. rs199930 has a submitted SNP (ss5411833) whose detail report shows genotypes of "C", "T" and "N". How can genotypes like these be possible, and why isn't this data displayed in the refSNP Cluster Report?**

Go to the Population Diversity section of the refSNP Cluster Report for rs199930. A member of the cluster, ss5411833, has a genotype from the population "CHMJ". When you click on the blue "CHMJ" text, you'll see a detailed description of the population sampled. This description indicates that the sampled was a "complete hydatidiform mole" (CHM) which happens to be haploid. dbSNP usually excludes haploid data from genotype frequency computation, but the FTP reports were already generated before the haploid genotype frequency data could be removed. Starting with next build, the FTP files will be corrected.(**7/28/06**)

**Why does dbSNP show each submitted SNP having its own genotype and frequency information? Isn't this is an over-representation of the number of unique SNPs with these particular attributes?**

Although several submitted SNPs are grouped within a single refSNP cluster, each submitted SNP of that cluster can have different sequence data. We capture genotypes on the submitted SNP level to preserve this underlying sequence data, which was used to design the assay for the SNP in question (probes, primers, etc.). When the information is available, Build 126 will have probe identifiers for HapMap genotypes that refer to an Entrez Probe record associated with the individual genotypes. (**5/23/06**)

## Genotype Detail (Genotype and Allele Frequency Report)

**In the "Genotype and Frequency Allele Report" for rs7903146, many of the alleles are red, indicating "inconsistent with at least one other submission". Which do I use?**

As you can see in the "Population Diversity" section of rs7903146's refSNP page, there are many discordant genotypes. I'm sorry to say that we at dbSNP cannot ascertain which is the correct genotype, since dbSNP only collects data and reports a discrepancy in order to alert the user. You must interpret the data yourself. We have no information regarding how or when the error or discrepancy was introduced: it could have been during the chip design, sample preparation and labeling, assay, or during data analysis.

To aid you in your analysis of the discordant data:

Check the submitters' websites to see if you can get any additional genotype information for this variation.

If you do not find helpful information in their websites, contact the submitters directly. By searching dbSNP for the handle, you will find their contact information.

Quite a few dbSNP users are encouraged to contact data submitters directly when they have specific questions about the submitters' data and have received helpful information.(**04/07/08**)

## Frequency

**Could you tell me the minor allele frequency for rs1050622?**

This SNP is monomorphic in a CEPH population, according to the frequency data that were submitted for this SNP.

You can see this by looking at the "Population Diversity" section of rs1050622's cluster report on the rs report. (**06/26/08**)

**Do you have a table that presents the minor allele frequencies for a list of SNPs for the same population found using different genotyping platforms?**

This information is available on in the population diversity section of the refSNP cluster report for each SNP.

**The refSNP report for rs2037101 shows this variation has a frequency of zero. Does this mean that this polymorphism is not confirmed by population data?**

The genotype for rs2037101 was submitted by Hapmap, and it indeed shows no variation frequency. If you click on the submitted SNP number (ss#) in the submitter records section of the refSNP report, you can see the details for ss2946401 (the original SNP for this refSNP). As you can see from this report, the original SNP was submitted by TCS. If you click on the link next to the word "Method" in the assay section of the submitted SNP (ss) report, you'll get a description of the method used to detect this SNP. In this case, the method used to discover the SNP was computational, and as such, it must be validated by genotypes, or by frequency, or by separate independent observations. As of this writing, that has yet to occur. (**11/03/06**)

**The refSNP report for rs2839858 doesn't provide the frequency of each allele and the population diversity section of the report shows that all individuals have the A/A genotype. How can this be a SNP if there are no other observed alleles?**

Frequency and genotype data is not required in order to submit SNP assay data, although dbSNP does encourage its submitters to provide frequency and genotype information when submitting a new SNP.

Sometimes SNPs are discovered using a computational algorithm, and in such a case, there are no frequency and genotype data available. Other submitters, however, may submit the frequency/genotype data for such a SNP at a later date. In the case of rs2839858, you can see in the "Population Diversity" section of the report

that the exemplar submitted SNP (ss4021194) genotype data was generated by HapMap. It is hard to say if this is truly a "Novariation" site. It might be a rare allele.

I encourage you to contact the submitter directly for more information. You can get the submitter's contact information by clicking on a submitted SNP number (ss#). For example, if you click on ss48294011, you'll go to the detail page for that SNP which contains the submitter information. Once on the detail page, click on the submitter's handle, in this case "SNP500CANCER" to get the contact information.

You can also click on the handle/ submitted SNP ID located in the "Submitter Records" section of the refSNP report to get information on the method of discovery directly from the submitter. For example, if you click on "SNP500Cancer ID|DIO2-05" for ss48294011, you'll go to the SNP500CANCER page , which contains further details about this particular submitted SNP.(**9/14/06**)

## Alleles

**The population diversity section of the rs11545130 cluster report shows an A-to-G change, but the integrated maps section shows a C nucleotide.**

You are correct that the cluster report shows the variation in two different orientations. The integrated maps section of the cluster report shows the variation in the reference sequence (NT_011520.11) orientation, while the population diversity section reports the frequency and allele data in refSNP orientation (the orientation of the cluster exemplar), which is why it was reported as A/G.

There are a number of reasons that we need to report RefSNP data based on RefSNP orientation defined by rs FASTA rather than based on reference sequence orientation:

- Sometimes a SNP maps to multiple reference sequences each of which have different orientations.
- Sometimes a SNP doesn't map to any genome position.
- Sometimes the NCBI reference sequence and the Celera or HuRef contigs are of different orientations.
- Between genome builds, contig sequences may change between different versions (although these changes becoming more rare as the contigs mature).

(**08/01/08**)

**How many times was the variation represented by rs8137714 observed? What population and how many chromosomes did you screen?**

Currently, rs8137714 does not have genotype or allele frequency information in dbSNP. To see how the submitted SNPs in this refSNP cluster were mined, do the following:

1. Go to the refSNP report for rs8137714 and look at the "Submitter Records" section
2. Click on either of the NCBI Assay ID numbers located on the left. This will take you to the submitted SNP Detail report for the selected SNP.
3. Scroll down to the Assay section of the Detail Report, and click on the assay "Method" listed there. This will take you to a description of the method used for finding the submitted SNP.

By following the above steps, you can see that both submitted SNPs in this cluster were computationally mined.

If you look at the Validation Summary" section located at the bottom of the refSNP report, you will see that this SNP is a "Double Hit" SNP. In other words, "all alleles have been observed in at least two chromosomes apiece." You can find more information on Double hit SNPs in the dbSNP FAQ Archive.

Some submitters will submit frequency information for existing refSNP numbers, so it is possible that dbSNP will get frequency information for this refSNP (rs) number in the future. (**11/26/07**)

**Two reports for rs3676330 claim that the variation for mouse strains C57BL/6J and 129S1/SVIMJ are T and A, respectively, whilst the third report claims the reverse. Which report is correct and why are the letters highlighted in green?**

The two submissions were on opposite strands, so they actually both agree and that is why they are colored green.

Expand the SNP detail section (by click on the blue plus sign) at the bottom of the page for strand orientation details. **(10/04/05)**

**How do I find the set of alleles that you used to instantiate a SNP allele sequence?**

We do not choose the alleles for a SNP. We include all alleles reported by submitters for a refSNP cluster. If a submitted SNP is on the reverse strand relative to the refSNP sequence, we reverse the alleles. For example, in the refSNP(rs) cluster rs268, rs268and ss268 have the allele A/G, while ss48420135 shows the C/T allele on the reverse strand. So the reported allele set for refSNP cluster rs268 is reported as A/G.

(**6/21/06**)

## The Validation Summary Section

**What exactly does it mean when a SNP is validated? Could you explain what validation is?**

In order for a RefSNP(rs) to be validated, at least one of its clustered submitted SNPs (ss) must either have been ascertained using a non-computational method or have frequency information associated with it.

When an ss is withdrawn from a validated rs cluster, and the withdrawn ss was the only ss in that cluster to have frequency information or to be ascertained using a non-computational method, then the rs cluster changes to"non-validated" status. For example, the submitter "SNP500CANCER" found all their SNPs using non-computational methods, and routinely withdrew SNPs during their quality control cycles. So when "SNP500CANCER" submitted a ss into dbSNP and it clustered into a non-validated rs, that rs became validated. When"SNP500CANCER" later withdrew the same ss, the rs cluster it was associatedwith lost its validation status.

You can also find information on variation validation by going to the dbSNP Handbook, and search for the text: "Validation" (scroll to the bottom of the page). You will find the following statement:

“dbSNP accepts individual assay records (ss numbers) without validation evidence. When possible, however, we try to distinguish high-quality validated data from unconfirmed (usually computational) variation reports. Assays validated directly by the submitter through the VALIDATION section show the type of evidence used to confirm the variation. Additionally, dbSNP will flag an assay as validated (Table 4) when we observe frequency or genotype data for the record.top link.” **(04/21/08)**

**Can you tell me what “Validation by HapMap” really means?**

"Validation by HapMap" in dbSNP simply means that a SNP was genotyped in HapMap (phase 1 & 2 over 270 samples, phase 3 over 1115 samples (not in dbSNP yet).

For some SNPs, HapMap found homozygous genotype results in the 270 samples. In these cases, the SNPs still have the "Validation by HapMap" flag, but will not have the "Validation by Frequency" flag ("Validation by Frequency" requires at least 2 minor alleles).

You should therefore look at "Validation by HapMap" in conjunction with "Validation by Frequency" to verify that the SNP’s minor allele has been observed at least twice. (**07/09/08**)

**dbSNP’s “Validation status description” states that one mode of validation is when “All alleles have been observed in at least two chromosomes apiece.” Do you mean chromosomes from different populations?**

There is a detailed discussion of the double hit validation in the "Build Process" section of the archive. (**10/25/07**)

**dbSNP's "Validation status description" states that one mode of validation is "Validation by Frequency or Genotype data". What is the difference between a validation by frequency or by geneotype?**

Validation by Frequency includes both population frequency data AND genotype data. In fact, the number of SNPs that have genotype data is bigger then the number of SNPs with only population frequency data. We compute frequency based on genotype data. (**10/25/07**)

**What is the meaning of the symbols that represent validation status in the RefSNP cluster report?**

Click on the underlined words "Validation Status" located in the Validation Summary section of the RefSNP cluster report (the validation summary section is located at the very bottom of the RefSNP cluster report). The underlined words are a link that will generate a pop-up window explaining the meaning of each validation status.

**How do I access the legend for the icons in the SNP validation section of the refSNP (rs) and submitted SNP (ss) reports?**

The link to the descriptions of the validation status icons are the underlined words "Validation Status", which appear in a blue font located just below the section heading.

**How does dbSNP define a double-hit SNP?**

A double-hit SNP is a SNP where all alleles have been observed in at least two chromosomes apiece.

**In the "VALIDATED" field, what does "by-cluster" mean?**

The cluster has two or more submissions, with one or more submissions assayed using a non-computational method. **(3/29/05)**

**In the "VALIDATED" field, what does "by-frequency" mean?**

At least one submitted SNP in the cluster has submitted frequency data. **(3/29/05)**

**How does dbSNP identify SNPs located in a particular gene, and how do I find out the number of people sequenced for each SNP? If too few people were sequenced, we will consider re-sequencing the entire gene.**

It looks like you want to know if a given SNP is "validated" or not, and that if a large percentage of SNPs on a gene are not validated, then you would re-sequence the whole gene.

To see how each SNP within a gene is sequenced may require lots of time to go through many details about the SNP in the dbSNP website. Using the SNP validation status is a more straightforward approach.

dbSNP sets the validation status of a SNP based on four conditions; you can see the details of these conditions by clicking on the validation status link. Several large labs, including that of the HapMap project, have genotyped many validated SNPs from dbSNP and found that most SNPs labeled validated in dbSNP are indeed variations.

Determining how each SNP is sequenced to assure yourself that the validation is adequate will take much more time. Most genes have SNPs submitted to dbSNP by different labs; therefore, you need to look at how each SNP is obtained in the publication and/or method section to see if they meet your criteria of trust. For example, look at human SNPs in the coding region of LPL.

Let's pick two rs numbers on LPL. rs1801177 did not have sample size data provided. From the refSNP page, you can see that this SNP was submitted by HGBASE, which collects data from publications. You can see more details by clicking in the publication section.

For rs1121923, follow the link of ss1455837, and you'll see it was sequenced on 221 individuals using a particular method.

# Sequence Data and Related Information

## Discrepancies between Sequences seen on the Web vs. those on the FTP site

**Why do web queries of rs939820, rs10205833, and rs7597158 return sequences that do not match the exemplar sequences for these SNPs found using a database query?**

When you refer to the sequences as "not matching", I assume that you are referring the fact that they don't match at the point of variation, since in rs939820, for example, the two flanking sequences are the same with the exception of the point of variation.

On the RefSNP (rs) page, we show the rs FASTA using the IUPAC code for variations, while on the submitted SNP (ss) page, we show the ss fasta using the submitted observed sequence. In most cases, all member ss of an rs cluster have the same allelic states in the same orientation, so the rs variation matches the ss exemplar variation. There are cases, however, where the rs variation does not match the ss exemplar variation. For example, if an ss exemplar has an A/G variation, and another ss from the cluster in the same orientation has an A/T variation, then the rs allele list will read A/G/T since it includes all member ss alleles. If you viewed the rs allele list converted into IUPAC code (remember the refSNP page shows the flanking sequence in IUPAC), it would show a D representing A or G or T.

Most of the submitted SNPs have the same variations in the rs clusters you mentioned, but one or two of the ss in each cluster have an extra allele. All of the ss in the rs939820 cluster have an A/G variation, with the exception of one ss that has an -/A/G variation. All the ss in the rs10205833 cluster have a C/G variation, with the exception of one ss that has a C/G/T variation. Most of the ss in the rs7597158 cluster have an A/G variation, while the ss exemplar has a -/G variation. In these cases, the refSNP page variation list includes all allelic states: for rs939820, instead of an R, it is N; For rs10205833, instead of an S, it is B that represents for C or G or T; for rs7597158, since the ss exemplar has a -/G variation, while most of the other ss in this cluster have an A/G variation, the refSNP FASTA shows an N at the variation point.

I will update dbSNP's variation representation for "mixed variation" clusters to include all the allele lists from all the submitted SNPs.(**8/18/06**)

## Sequence Orientation

**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**)

**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**)

**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.

**I have found a refSNP in dbSNP with its flanking sequence in reverse orientation (anti-sense orientation) for the LIG1 gene. Kindly update your database.**

rs11879148 is in reverse orientation with respect to mRNA NM_000234 for the LIGI gene because the flanking sequence of a submitted SNP for that cluster was used as the refSNP (rs) cluster flanking sequence, so the orientation of the flanking sequence for a refSNP cluster doesn't depend on the orientation of a contig sequence or mRNA sequence.

The mRNA orientation column in the "GeneView" section of the refSNP page for rs11879148 shows that this SNP is in reverse orientation with respect to the mRNA for LIG1. To insure that a refSNP's orientation is stable throughout various builds, we do not change the refSNP flanking sequence. We also apply this rule to those SNPs that map to multiple positions or do not map to genome at all.(**6/29/06**)

**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**)

**rs4897909 and rs4788229 are 4 base pairs apart, but if I superimpose the flanking sequences of both, the "K" variation of rs4897909 corresponds to an "A" base on the flanks of rs4788229.**

The fact that rs4897909's G/T(K) aligns with rs4788229 base "A" in the same orientation does seem curious. These two rs numbers are both from the same computational SNP discovery program (SSAHA — you will find references to this program by reviewing the publications associated with these SNPs), and there are no other submitted SNP clusters that map to the same position. Also, there is no population frequency or individual genotype validation information available for these two SNPs.

Based on above information, I'm guessing that the "K" erroneously aligns to "A" in the same orientation for the following reasons:

First, the NCBI build has progressed to 36, whereas these SNPs were discovered on build 31. It might be that the build 31 contig upon which the SNP comparison was based contains errors. An example supporting this supposition: rs4897909 was based on a single submitted SNP (ss) from SSAHA: WI_SSAHASNP|NT_025920.10_372470, but searching NCBI for NT_025920 shows that this contig has been removed from the current genome build.

Second, dbSNP sets the SNP validation status based on frequency/genotype information or multiple submissions from non-computational methods. Roughly half of the SNPs in dbSNP are validated. At the present time, these two SNPs are not validated, and therefore can be considered suspect. (**9/5/06**)

### Sequence Notation

**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**)

**In SNP flank sequence, I find that some bases are capital letters, while others are in lowercase (small) lettering. What is the difference between the two?**

Sequence in lowercase (small letters) 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)**

**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)**

### Determining Distance of Submitted Flanking Sequence for Variation

**How do I determine the distance up and downstream that the sequence for rs3208856 extends?**

The sequence submitted for this rs number is located in the FASTA section of the refSNP cluster report for rs3208856. The FASTA data show that rs3208856 has 50 bp on each side of the variation. Go to the Integrated Maps section of the RefSNP report, and click on the rs number link located in the NCBI Sequence Viewer field to see more sequence surrounding this SNP.

## Phenotypic Data

**Does dbSNP curate phenotypic information to include in its reports?**

dbSNP does not curate phenotypic information, but we hope to include some in the future. We do have some SNPs mapped to allelic variants in the Online Mendelian Inheritance in Man (OMIM) database.

## The Difference between refSNP (rs) and Submitted SNP (ss) numbers

**I'm new to dbSNP and am a little confused about the definitions of ss numbers and rs numbers. What is the difference between the two?**

An ss number is the unique ID number assigned to each submitted SNP. Once the ss number is assigned, we align the flanking sequence of each submitted SNP (ss) to the contig. If several ss numbers map to the same position on the contig, we cluster them together and call the cluster a "reference SNP cluster", or a "refSNP". Each refSNP is given a unique rs ID number, and the flanking sequence for each rs number is always the longest flanking sequence found among all the ss flanking sequences of a particular refSNP (rs) cluster.

If there is only one ss number that maps to a specific position, then that refSNP cluster will contain only one ss number. In such a case, the ss flanking sequence and rs flanking sequence are the same. The flanking sequence may change later, however, if another ss number joins the cluster that contains a longer flanking sequence.

You many find this documentation helpful in explaining dbSNP.

# Locating Specific Data in a RefSNP Report

## Finding Amino Acid Changes Resulting from a Variation

**On the refSNP page, where do I find information about amino acid changes resulting from the variation?**

When a SNP causes an amino acid change, you can see the changes in the Gene View section of the refSNP report (example). The Gene View section is located about half way down the refSNP report page. (**1/13/06**)

## Finding a Graphical Display of SNPs in a Gene Region

**Where is the resource that allows you to actually see in a graphic how many SNPs are known to lie within a particular gene, and their physical coordinates?**

The resource you refer to is the sequence viewer graphical display of a gene region on the human genome. We had to discontinue this resource because the NCBI sequence viewer itself fails when the sequence under consideration is very large (contig sized).

You can still get a sequence view of the mRNA record using the seqview button on the Entrez SNP display, but this won't show intronic SNPs. There may be other inconsistencies even among coding SNPs since the position of SNPs on mRNAs is taken from a different mapping pipeline then the one used to compute genomic locations.

For now, you can search Entrez SNP using the gene symbol or gene id and click the 'GeneView' icon on any of the displayed rs records:

1. Go to Entrez SNP and enter the search terms "LPL AND Human" (without the quotation marks) into the text box at the top of the page, and click on The "Go" button.
2. You will get a page back that contains 158 results to your query.

You can also use Entrez SNP to get a rough text representation of a SNP position in a gene by entering the gene symbol in the search box at the top of the page. When you get the result page, select "chromosome report" from the "display" drop-down menu and "Chromosome Base Position" from the "sort by" drop-down menu. **(6/3/05)**

**Are there any refSNP fields that indicate which SNP allele matches the reference genome sequence upon which the SNP was mapped?**

We maintain the contig variant in one of our database tables, called SNPContigLoc, which is located in the organism_data directory of your organism's database. Once you locate the organism_data directory (the previous link is to the human_9606 directory as an example), download SNPContigLoc.bcp.gz. The first column of this table is snp_id. The last column of this table contains the contig form of the variation (in SNP orientation rather than in contig orientation). When the orientation in the next to the last column of this table is set to one, you'll have to reverse complement if you want the contig orientation.

## Finding Contact Information for a Submitter

**Where do I find contact information for the submitter of rs12638783, as well as more detailed information about this refSNP?**

You can get reports for a particular SNP by going to the dbSNP home page and entering "rs12638783" in the text box located in the "Search by IDs" section of the page, and then pressing the "search" button.

There is a Reference SNP (rs) report as well as a submitted SNP (ss) detail report and a ss detail report in submission format available for rs12638783. To get contact information for the submitter, click on the submitter handle in either of the ss reports. Please note that no frequency data was submitted for this SNP. **(4/13/05)**

**How do I determine who submitted rs11544612?**

Go to the rs11544612 refSNP page and scroll down to the "Submitter Records for this refSNP Cluster" section. Click on the text that reads: "ss16240609" (the only submitted SNP for this cluster) to go to the "Submitted SNP Detail" page for this ss number. Once you are on the detail page, click on the text for the submitter handle, which in this case is "CGAP-GAI", to go to the "SNP Submitter Contact Detail" page, which contains all full name of the submitting individual or institution along with their contact information. (**9/7/06**)

## Finding SNP Citations (SNP publication details)

**How can I find the references for ss42780946 & ss460400? I've looked in what I thought to be obvious places, but can't find them in dbSNP.**

Here is how you find the references:

1. Go to the dbSNP home page, and enter ss42780946 in the text box located in the "Search by ID on all Assemblies" section located just below the announcement section near the top of the page, and select "NCBI Assay ID(ss#)" from the drop down menu to the right of the text box where you entered ss42780946. Click the "search" button.
2. The resulting submitted SNP report will appear. Scroll down this report until you get to the "Individual Genotype Batch" section where you'll see "View citation details". Click on the numeral "1" link just to the right of these words.
3. The resulting publication detail page does not indicate the actual reference, but if you click on the number that follows the word "PMID" near the top of the page, you will get to the actual publication reference.

I'm sorry the link to the reference is not clear, but I will mention it to the SNP development team, and perhaps they can work out something that will make it easier to see. (**8/8/07**)

# Definitions for Terminology Used in the refSNP Report

## TER

**Is "Ter" an abbreviation for a termination codon?**

Yes. "Ter" is an abbreviation for a termination codon. **(9/30/05)**

**What does TER[*] for a non-synonymous coding change mean?**

TER[*] means that the variation changed the codon to a termination codon (Ter or *), which causes premature termination of the protein. (**10/25/06**)

## cSNP

**What does cSNP mean?**

A cSNP is a SNP located in the coding region (exon) of a gene.

## Complement

**Some synonymous SNPs in dbSNP are annotated on mRNA as "complement", and are not shown on the protein. What does "complement" mean, and why is annotation different for dbSNP and RefSeq?**

"Complement" means that the SNP mapped to the complementary strand of the mRNA. The SNP will map to the compliment if the mRNA is in the reverse orientation with respect to the genome and the SNP is maps in the forward orientation.

Example:

variation complement(1601)

/gene="IFIH1"

/gene_synonym="Hlcd; IDDM19; MDA-5; MDA5; MGC133047"

/replace="a"

/replace="g"

/db_xref="dbSNP:10930046"

In the above example, the mRNA is on the reverse orientation with respect to the genome, and the SNP (i.e. rs10930046 [C/T])is on the forward orientation. You can see this in the GeneView section of the refSNP cluster page for this SNP. Synonymous SNPs that do not change the amino acid are not reported on the protein.(**06/19/09**)

## Contig Reference

**What does “contig reference mean?**

“Contig reference” refers to an allele on a contig at a particular SNP position. For example, if a synonymous or non-synonymous SNP at a particular contig location has the variation of “C/T” and the “contig reference for that SNP is “T”, then the SNP would be “T” at that SNP location on the contig.

## Double-hit SNP

**How does dbSNP define a double-hit SNP?**

A double-hit SNP is a SNP where all alleles have been observed in at least two chromosomes apiece.

## Population

**What is the definition of the population described in a SNP report?**

A "population" is submitted and defined by a submitter, which means that the submitter provides a population ID (an abbreviation) as well as a description of the population, therefore, “populations” are not curated by dbSNP.

Please note that genotype data also have populations, but the "population with frequency data" designation is only for populations that have allele frequency submissions, and these do not include genotype submissions.

You can see how major contributors to dbSNP define their populations by going to the “Search/View Population Detail” page and type in the submitter handles "Perelgen" and "CSHL-HapMap" in the search text box located in the grey search section and clicking the “Search” box directly below. **(3/9/05)**

### Synonymous vs Non-Synonymous

**What does it mean when an SNP is labeled "synonymous" or "non-synonymous"?**

The terms "synonymous" and "non-synonymous" are used for SNPs that are in predicted protein coding regions (i.e., exons of genes). Synonymous SNPs are those SNPs that have different alleles that encode for the same amino acid. Non-synonymous SNPs are SNPs that have different alleles that encode different amino acids. For further details, I recommend querying for "genetic code" or "protein translation" at the NCBI books website.

### Sample sizes

**What is the difference between assay sample size and population data sample size?**

The assay size is the number of chromosomes examined to ascertain or detect the SNP, while the population sample size is the number used in determining the frequency of the SNP. Sometimes the sample size will be the same in both cases if the same group detected and determined a SNP frequency using the same population. **(9/1/05)**

## Submitted SNP (ss) Detail Reports

**The submitted SNP page contains the term "frequency of minor variant". What does it mean?**

The frequency of the minor variant is the minor allele frequencies for heterozygous and homozygous SNPs with reference to the frequency of all alleles at a particular SNP location.

For example, suppose a SNP is genotyped on 50 individuals, with A/A=45 and G/G=5, the minor allele is therefore G with a frequency of 0.1. If the genotype count for the same 50 individuals was reported as A/A=43, A/G=4, and G/G=3, the minor allele would therefore be G, with a minor allele frequency still at 0.1. (**5/1/06**)

**In the comment section of the report for a submitted SNP(ss) that I was examining, it said "quality: 51". What does this mean?**

The text provided in the Comment section is annotation information provided by the submitter, so the "quality" tag is not a dbSNP-defined field. I suspect "quality" as referred to by the submitter is most likely the Phred quality value, as described the submitter assay method.

**I'm new to dbSNP and am a little confused about the definitions of ss numbers and rs numbers. What is the difference between the two?**

An ss number is the unique ID number assigned to each submitted SNP. Once the ss number is assigned, we align the flanking sequence of each submitted SNP (ss) to the contig. If several ss numbers map to the same position on the contig, we cluster them together and call the cluster a "reference SNP cluster", or a "refSNP". Each refSNP is given a unique rs ID number, and the flanking sequence for each rs number is always the longest flanking sequence found among all the ss flanking sequences of a particular refSNP (rs) cluster.

If there is only one ss number that maps to a specific position, then that refSNP cluster will contain only one ss number. In such a case, the ss flanking sequence and rs flanking sequence are the same. The flanking sequence may change later, however, if another ss number joins the cluster that contains a longer flanking sequence.

You many find this documentation helpful in explaining dbSNP.

**C and G are listed as alleles in the "Summary of Genotypes" section of the detail report for ss15377600, yet the "Allele" section in the same report shows that the observed alleles are -/T.**

One of the idiosyncrasies of dbSNP is that genotype & frequency data need to be linked to one of the submitted-SNP(ss) records within a refSNP(rs) cluster — specifically the ss exemplar for that cluster (see FAQ regarding ss exemplars for a refSNP) — because a refSNP will sometimes merge away. Linking genotype & frequency data to the ss exemplar becomes a problem when different submitted SNPs contribute different variations to the refSNP cluster. This is the problem with the submitted SNP (ss15377600) you mention in your question. In this case, ss15377600 happens to be the exemplar for the refSNP cluster, and is an in/del variation, while all other ss in the rs2070922 cluster are true SNPs and contribute the allele frequencies you found in the "Summary of Genotypes" section of the report.

The SNPdev team is thinking about separating refSNP clusters if the exemplar submitted SNPs within that cluster is of a different class from the other submitted SNPs in the cluster (such as indel vs. true SNP, as in this example). I will try to determine how many ss exemplars do not have the alleles reported in their refSNP genotypes. In the meantime, please look at the refSNP allele list to see if a submitted SNP genotype allele is valid or not. **(2/28/05)**

**What are the "Y" and "N" designations for SNP genotypes in dbSNP? I was looking at the submitted SNP (ss) report for ss538, and found the designation N/N…**

"N" is the designation for "tested, but results indeterminate".

"Y" is the designation for genotypes from males for SNPs that map to the X chromosomes. (**01/11/08**)

## VarView

**There is a link to "VarView" in the Allele column of the cluster report for a number of refSNPs I'm looking at. What exactly is "VarView"?**

"VarView" (short for "Variation Viewer") icons or links result in a gene-specific display named for the variation view of the gene in question: "[Gene Symbol]" + "Variation Viewer" (e.g. MECP2 Variation Viewer).

VarView is an improved alternative to dbSNP's GeneView in that it contains more intuitive packaging of the data found in dbSNP records (e.g. HGVS names, numbers of observations, clinical associations, links to OMIM and Locus-specific databases (LSDB), citations, etc.).

Currently, a VarView link or icon appears in a variation record only if the variant was submitted with clinical association(s), and if the gene in question has a RefSeqGene record.(**09/22/08**)

**How do you associate dbSNP entries with PubMed articles in VarView?**

The publication links in VarView come from two different sources:

- PubMed IDs are submitted within a dbSNP submission.
- dbSNP has begun mining Pubmed for dbSNP links, which will be formally added to dbSNP soon. (**05/08/08**)

**The "Variation Class" column in the "Observed Variation" section of the Variation Viewer is showing a new variation class called SNC. What exactly is the SNC variation class?**

SNC represents "Single Nucleotide Change". As there was some confusion fostered by our use of SNP (Single Nucleotide Polymorphism) for variants that were not polymorphic, we have moved to this descriptor. (**12/22/08**)

# ASN.1. Reports

**I noticed in the human chromosome 22 ASN.1 flat files that a couple of records have the same contig, but their CTG start and end are different.**

This will happen when some SNPs map to multiple locations on the genome and chromosomes.(**10/23/08**)

**I am looking at the ASN.1 flat file of a SNP and find that there are two genomic locations of a single SNP. Which is correct?**

Here are the ASN.1 Flat file lines that were causing you difficulty:

```
CTG | assembly=Celera | chr=X | chr-pos=1068500 | NW_927672.1 |
ctg-start=133190 | ctg-end=133191 | loctype=3 | orient=+
CTG | assembly=reference | chr=Y | chr-pos=12606579 | NT_011875.11 |
ctg-start=298001 | ctg-end=298002 | loctype=3 | orient=+
```

In the above flatfile lines, we see that this SNP is mapped to both the NCBI reference assembly (see above:"assembly=reference") and the Celera assembly. (see above:"assembly=Celera"). So both positions are correct so long as you specify that the variation is on either the reference assembly or on the Celera assembly.

Please note that CTG is an abbreviation for "Contig", and there is a detailed explanation of Loctype available online at the dbSNP website. (**12/28/07**)

**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 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.

**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)**

**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**)

# Chromosome Reports

**Can you define the columns in the "chromosome reports" format?**

For future reference, you can find column definitions for chromosome reports in the README file on the ftp site. Below you will the column definitions for chromosome reports which I have copied from the FTP README:

CHROMOSOME REPORTS

---

The Chromosome Reports format provides an ordered list of RefSNPs in approximate chromosome coordinates (the same coordinate system used for the NCBI genome MapViewer); it is a small file to download, and contains a great deal of information about each SNP.

The lines of the Chromosome Report format give the following information for a single RefSNP in tab-delimited columns:

| Column | Data |
|---|---|
| 1 | RefSNP id (rs#) |
| 2 | mapweight where:<br>1 = Unmapped<br>2 = Mapped to single position in genome<br>3 = Mapped to 2 positions on a single chromosome<br>4 = Mapped to 3-10 positions in genome (possible paralog hits)<br>5 = Mapped to >10 positions in genome |
| 3 | snp_type where:<br>0 = Not withdrawn.<br>1 = Withdrawn. There are several reasons for withdrawn, the withdrawn status is fully defined in the asn1, flatfile, and XML descriptions of the RefSNP. See /specs/docsum_2005.asn for a full definition of snp-type values. |
| 4 | Total number of chromosomes hit by this RefSNP during mapping |
| 5 | Total number of contigs hit by this RefSNP during mapping |
| 6 | Total number of hits to genome by this RefSNP during mapping |
| 7 | Chromosome for this hit to genome |
| 8 | Contig accession for this hit to genome |
| 9 | Version number of contig accession for this hit to genome |
| 10 | Contig ID for this hit to genome |
| 11 | Position of RefSNP in contig coordinates |
| 12 | Position of RefSNP in chromosome coordinates (used to order report)<br>Locations are specified in NCBI sequence location convention where:<br>x, a single number, indicates a feature at base position x<br>x..y, denotes a feature that spans from x to y inclusive.<br>x^y, denotes a feature that is inserted between bases x and y |
| 13 | Genes at this same position on the chromosome |
| 14 | Average heterozygosity of this RefSNP |
| 15 | Standard error of average heterozygosity |
| 16 | Maximum reported probability that RefSNP is real. (For computationally-predicted submissions) |

*Table continued from previous page.*

| Column | Data |
|---|---|
| 17 | Validated status<br>0 = No validation information<br>1 = Cluster has 2+ submissions, with 1+ submission assayed with a non-computational method<br>2 = At least one subsnp in cluster has frequency data submitted<br>3 = Non-computational method in cluster and frequency data present<br>4 = At lease one subsnp in cluster has been experimentally validated by submitter<br>for other validation status values, please see:<br><ahref="ftp://ftp.ncbi.nlm.nih.gov/snp/database/<br>organism_shared_data/SnpValidationCode<br>.bcp.gz">ftp://ftp.ncbi.nlm.nih.gov/snp/database /organism_shared_data/Snp Validation Code<br>.bcp.gz</a> |
| 18 | Genotypes available in dbSNP for this RefSNP<br>1 = yes<br>0 = no |
| 19 | Linkout available to submitter website for further data on the RefSNP<br>1 = yes<br>0 = no |
| 20 | dbSNP build ID when the refSNP was first created (i.e. the create date) |
| 21 | dbSNP build ID of the most recent change to the refSNP cluster. The date of the change is represented by the build ID which has an approximate date/time associated with it. (see the dbSNP Build history). |
| 22 | Mapped to a reference or alternate (e.g. Celera) assembly |

Also included within the chr_rpt file are two additional files:

multi/ contains a list (in chromosome report format) of SNPs that hit multiple chromosomes in the genome

noton/ contains a list (in chromosome report format) of SNPs that didn't hit any chromosomes in the genome

(**07/22/08**)

**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)

**Is there a difference between the Chr_MT and the Chr_Multi files that are located in the chr_rpts and XML directories?**

The two are different but both are based on the reference assembly:

Chr_MT: is a file that contains SNPs that fall within mitochondrial DNA.

Chr_Multi: is a file that contains SNPs with a map weight greater than 2, where:

- weight 1 = SNPs that align at exactly one locus
- weight 2 = SNPs that align at two loci on same chromosome
- weight 3 = SNPs that align between 3 and 10 loci (hits do not have to be on same chromosome)
- weight 10= SNPs that align at greater than 10 loci (hits do not have to be on same chromosome)

(**02/08/08**)

# FASTA Reports

**Is there any way of downloading each Human chromosome in a fasta file with all synonym SNPs integrated in the sequence in IUPAC code?**

dbSNP doesn't have fasta reports containing IUPAC code for the variations.(**06/03/09**)

**I am submitting sequence to Illumina for Golden Gate assay oligo design, which requires a Microsoft Excel file in csv format, where each FASTA sequence is in a single cell of the file. Can I retrieve data from dbSNP in this format?**

Both flanks and the alleles need to be in the same cell of an excel file. An adjacent cell in the same row would have the rsID number.

The data you need can be obtained from dbSNP's current FTP reports.

For example, you can look at the FASTA report for dog

Please note that the sequences are directly from submissions. You may want to verify the sequences during your oligo design.(**06/02/08**)

**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.

# FLATFILE Reports

**We have a program that annotates mRNA sequence with tags for SNPs or in/del (insertion/deletion) mutations. We are having a problem running the program with the in/del mutations because the in/del refSNP records in the dbSNP flat files have changed from having "^" or ".." marks in the record (our program looks for the ".." marks), to not having these marks at all. What do we do?**

The change you have noticed is described on the Column Description for table: SNPContigLoc page. Look in the phys_pos column.

Where your program is looking for "..", try updating it so that it checks for ctg_start and ctg_end. ".." is used for ctg_start and ctg_end, which are not the same. Here's an example of the change:

The records used to look like this:

chr-pos=56091347..56091351 ( I omitted other fields here ) ctg-start=23667725 | ctg-end=23667729

Now, the above will be reported as:

chr-pos=56091347| ctg-start=23667725 | ctg-end=23667729

If your program could do the following:

If ( ctg-end - ctg-start ) > 0,

Then the chr-pos-range could be represented as:

(chr-pos),(chr-pos+ctg-end - ctg-start)

(**7/18/06**)

# Genotype Reports

**After reviewing dbSNP genotype reports for SNPs located on chromosome X in the GLA gene, I noticed that you consider a male as having two alleles. Shouldn't the statistical data be related to the number of chromosomes analyzed?**

Thank you for pointing out this problem in our display of hemizygous genotypes.

We agree with you that genotypes of male samples typed on X chromosome markers should have a single allele reported. Unfortunately some of our major data contributors (i.e. HapMap) continue to use the convention of reporting two alleles.

Our current recommendation to users submitting new genotypes for non-pseudo-autosomal X chromosome marker on male sample, is to encode the genotype as A/Y, where A is the observed allele, and Y is a place holder denoting the hemizygous state of the genotype.

We are working on creating the processing rules to fix the allele frequency and genotype counts given to these genotypes, as well as display issues with the genotype instances themselves. These changes will be incorporated as soon as possible.(**2/20/07**)

# XML Reports

## 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

**I have encountered single letter genotypes for SNPs in XML files that are not located on chromosome Y. rs199930 has a submitted SNP (ss5411833) whose detail report shows genotypes of "C", "T" and "N". How can genotypes like these be possible, and why isn't this data displayed in the refSNP Cluster Report?**

Go to the Population Diversity section of the refSNP Cluster Report for rs199930. A member of the cluster, ss5411833, has a genotype from the population "CHMJ". When you click on the blue "CHMJ" text, you'll see a detailed description of the population sampled. This description indicates that the sampled was a "complete hydatidiform mole" (CHM) which happens to be haploid. dbSNP usually excludes haploid data from genotype frequency computation, but the FTP reports were already generated before the haploid genotype frequency data could be removed. Starting with next build, the FTP files will be corrected.(**7/28/06**)

**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.

## Frequency Data

**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**)

## 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

## 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**)

**How do you assign strand orientation to a refSNP (rs) in the MapLoc element (from the Primary Sequence element), and how is the "orient" attribute determined?**

The "orient" value is determined by mapping SNP flanking sequences to a contig sequence using BLAST.

**Please note:** a refSNP(rs) flanking sequence will never change orientation even if submitted SNPs in the opposite orientation are assigned to the refSNP (rs) cluster. Also, during a build, refSNP flanking sequences are BLASTed against contigs. If an rs maps to different contigs in each of two different builds, then the strand (or "orient" attribute) of the two different contigs to which the rs maps are in opposite orientation.

Here is an XML example of the above:

```
<PrimarySequence>
<PrimarySequence_dbSnpBuild>126</PrimarySequence_dbSnpBuild>
<PrimarySequence_gi>8077580</PrimarySequence_gi>
<PrimarySequence_source value="blastmb"/>

<PrimarySequence_accession>AC027456</PrimarySequence_accession>
<PrimarySequence_mapLoc>
 <MapLoc>
 <MapLoc_asnFrom>189600</MapLoc_asnFrom> <MapLoc_asnTo>189600</MapLoc_asnTo>
 <MapLoc_locType value="exact"/>
 <MapLoc_alnQuality>0.999994</MapLoc_alnQuality>
 <MapLoc_orient value="forward"/>
```

```
<MapLoc_leftFlankNeighborPos>628</MapLoc_leftFlankNeighborPos>
<MapLoc_rightFlankNeighborPos>630</MapLoc_rightFlankNeighborPos>
<MapLoc_leftContigNeighborPos>189599</MapLoc_leftContigNeighborPos>
<MapLoc_rightContigNeighborPos>189601</MapLoc_rightContigNeighborPos>
<MapLoc_numberOfMismatches>1</MapLoc_numberOfMismatches>
<MapLoc_numberOfDeletions>0</MapLoc_numberOfDeletions>
<MapLoc_numberOfInsertions>0</MapLoc_numberOfInsertions>
</MapLoc>
</PrimarySequence_mapLoc>
 </PrimarySequence>
```

(8/17/06)

## 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**)

## 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.

## Chr_MT and Chr_multi

**Is there a difference between the Chr_MT and the Chr_Multi files that are located in the chr_rpts and XML directories?**

The two are different but both are based on the reference assembly:

Chr_MT: is a file that contains SNPs that fall within mitochondrial DNA.

Chr_Multi: is a file that contains SNPs with a map weight greater than 2, where:

- weight 1 = SNPs that align at exactly one locus
- weight 2 = SNPs that align at two loci on same chromosome
- weight 3 = SNPs that align between 3 and 10 loci (hits do not have to be on same chromosome)
- weight 10= SNPs that align at greater than 10 loci (hits do not have to be on same chromosome)

(02/08/08)