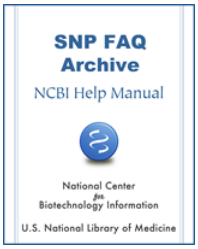
U.S. National Library of Medicine
National Center for Biotechnology Information

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



Interpreting Discrepancies in refSNP Reports

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

The "Created/Updated" Field

There are two submissions for rs4328 in dbSNP (build 36 and build 116), yet the "Created/Updated in build" flag in the cluster report says 36/126, indicating that there should be a submission from 126. Then I noticed in b126, that there are submissions from build 121 and 126 now missing in build127…

The "Created/Updated in build" flag was meant to show

1. The build when a refSNP was created, and
2. The last time the cluster had either added or deleted (submitter withdrawal) a member submitted SNP (ss)).

In this case, as you have observed, one submitted SNP (ss49841044) was withdrawn by the submitter in build 126. That was why the "Updated in build" is 126. I can see though, that since the refSNP page does not list submitted SNP numbers that were withdrawn, the "Updated in build" incrementing due to a withdrawal might be confusing. We will think about how to present this information more clearly. (**7/02/07**)

Allele section

rs1805010 used to be A/G, but now dbSNP lists it as A/C/G/T even though the cluster and population diversity data still reports only A and G.

The refSNP page for rs1805010 shows a large number of submitted SNPs(ss) in this cluster. With the exception of one ss (ss5105842 from submitter CBTMIGBGHOSH|BGID-4), all of the ss in the cluster reported an A/G allele in the same orientation as the refSNP (rs).

Although ss5105842 also reported "A/G" alleles, it gave the flanking sequence on the reverse strand. As dbSNP reports alleles exactly as they are submitted to us, if one submission provides the alleles in the wrong orientation, it will mess up the allele list for the entire refSNP cluster.

Looking at all the evidence, including the genotype data and the contig alleles, I think CBTMIGBGHOSH likely made a mistake when they submitted ss5105842. I will withdraw ss5105842 so it will not cause the extra allele issue. (**06/26/09**)

rs61734154 maps to the reference sequence in fwd orientation and has an "A" allele, whereas the submitted allele on the fwd strand is C/T.

I confirm that what you have described is correct:

rs61734154 does map to reference assembly in the plus orientation and that at the SNP position the reference assembly shows allele "A", whereas the submitted SNP(ss86258677) that comprises rs61734154's refSNP cluster has allele "C/T".

There are two possible explanations for this discrepancy:

- The submitter submitted the allele on the opposite strand of the flanking sequence.
- At the SNP position, it is indeed possible to have three different alleles: C,T and A. This is extremely rare (although it does happen often at some highly variant regions like MHC).

When there is no genotype data to further validate this SNP, it is hard to pinpoint the exact reason for this discrepancy.

To see if perhaps this discrepancy was the result of the first explanation mentioned above, I checked other SNP (ss86236922) that was submitted in the same batch to see if there is a systematic strand issue.

ss86236922 was submitted in Jan 2008, aligns to the reverse strand of rs603893, and has "A/G". The rs603893 cluster, however, contains a number of submitted SNPs where "C/T" is on the reverse strand while "A/G" is on the positive strand. Since rs603893 maps to reference assembly on reverse strand, and reference sequence has a "C" at the SNP position, rs603893 having "A/G" makes sense.

Based on this information, it is clear that ss86236922 seems to also have the allele on the wrong strand.

We will contact the submitter for this batch to ask them double check the allele strands. (**08/20/08**)

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**)

The cluster report for rs11466345 indicates that the allele is "A/G" and the ancestral allele is G, but some of the submitted SNPs are "C". Why?

The reason you see the "C' is that these submitted SNPs hit on the reverse orientation. You can see that the cluster exemplar hits in the reverse orientation when you look at the "Integrated Maps" section of the cluster report, which shows a "minus" in the "Hit orientation" column. By the way, the integrated maps section shows that the Reference and Celera assemblies have different alleles for this SNP: C and T. **(03/27/08)**

According to the literature, the alleles of rs7528684 (FCRL3) are T/C, but dbSNP shows they are A/G.

Looking at the GeneView section of the refSNP cluster report for rs7528684, you can see that rs7528684 is located on the reverse strand of the mRNA, which might be why the literature lists the alleles as T/C while dbSNP lists them as A/G. The submitter records section of the refSNP cluster report for rs7528684 shows 8 submissions of this SNP, each of which reports the alleles as A/G.

I did find a paper by Thabet, et al. that lists allele for this variation as T/C.

We encourage you to contact the authors of the articles you are interested in and ask why the alleles they cite are the opposite of those submitted to dbSNP. We would appreciate it if you could let us know what the authors' response is. (**03/17/08**)

In dbSNP, rs1548655 is G/T, while in Ensembl, it is A/C. According to Ensembl, they display only the fwd strand on their webpage. Which SNP is on which strand?

In Ensembl, the sequence of rs1548655 aligns to the strand opposite that of rs1548655 in the dbSNP website (you can use blast2seq to check the alignment yourself). That is why dbSNP reports the snp as having "G/T" alleles and Ensembl reports the SNP as having "A/C" alleles.

You mentioned that "According to Ensembl, they display only the fwd strand on their webpage." Is it possible that Ensembl's "fwd" means "fwd" with regard to genome orientation? In this case, rs1548655 on dbSNP maps to the minus strand of genome. So if Ensembl always show their SNPs in genome orientation, they would reverse the sequences and alleles for rs1548655.

I checked a few other SNPs (e.g. rs5, rs15) in dbSNP that map to the minus strand of genome, and sure enough, the SNP's sequences and alleles are reversed in Ensembl. So it looks like Ensembl's sequences are always the same as the genome. This is just from my limited observations, however, and I will have to confirm this with Ensembl. (**02/13/08**)

Does rs4818 I have 3 alleles or 2? The frequency does not show the T allele.

rs4818 has 28 submitted SNPs(ss). 27 of the submitted SNPs reported were C/G, while a single submitted SNP (ss16240701) from CGAP-GAI was G/T.

If you look at the detail page for ss16240701, you can see that this SNP was computationally mined from public EST data. Given that ss16240701 was computationally mined, and that it was the only ss out of 27 to report a genotype of G/T, I would venture to guess that the "T" allele is not real. (**02/14/08**)

Why does dbSNP refer to an A/G variation in the sequence and ancestral description of rs2476601, when several reference publications indicate that it is a 1858C>T polymorphism?

This discrepancy is the result of the opposing alignment orientation of the SNP(rs) flank and mRNA to the genome contig, and affects the presentation of the data only.

rs2476601 aligns in the same orientation as the genome contig, and the contig, at this variation site, has the allele "A" as noted in the "Allele" section at the top of the refSNP Cluster report for rs2476601.

The mRNA, however, aligns to the genome contig in reverse orientation. The mRNA's reverse orientation is noted in the Geneview section of the refSNP Cluster report for rs2476601 under the column "mRNA Orientation", and the dbSNP allele column notes the ":C" allele as present. The variation allele is indeed a "T to C" change when the variation is aligned in the mRNA's orientation. (**11/07/07**)

The alleles for rs1611430 are reported A/G/T, yet the variation class is SNP. I thought this classification was reserved for biallelic SNPs, and that tri- and multi-alleles were classified as MNPs.

MNPs in dbSNP are variations that have alleles with multiple nucleotides like AT/GT.(**6/8/07**)

The Submitter Records Section

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**)

Geneview Section

The rs1042714 report shows "G" as the contig reference allele (GeneView section) and "C" as the missense allele, but the pop diversity section shows "C" as more frequent.

The allele that is labeled "contig reference" is the allele on the contig at the SNP position. Contigs are assembled from sequences produced in the human genome project. The major genome assemblies include NCBI reference, Celera and the HuRef assembly.

Contig alleles should not be confused as wild types. It is entirely possible that a contig happens to have the minor allele, as you have observed in this case.

The term "missense" is used to flag an amino acid change caused from a SNP allele change relative to the contig. Missense refers to the type of non-synonymous codon change that is not "nonsense"(or stop codon), nor "frameshift".

We only know which allele is major and which allele is minor in a specific population when genotype or frequency data is available as is the case for rs1042714. So you are looking at the right information.(**08/22/08**)

Why does HuRef conflict with the NCBI reference assembly more than Celera (ex: HuRef reports extra gene near rs3168 that NCBI & Celera don't report)?

My guess is that the conflicts you have observed may be due to the fact that the reference and Celera assemblies are "consensus" assemblies that capture commonalities from pooled samples, whereas HuRef is from a single diploid assembly that may have individual specific differences with the reference assembly. (**08/07/08**)

I have been looking at the Geneview pages for a number of SNPs in OFCC1 and was wondering what "intron_0" means?

In the dbSNP Geneview page, we provide the exon numbers starting with "1" from 5' to 3' on the mRNA. The intron region immediately following exon 1 is numbered as "intron 1". In the OFCC1 gene, there are two UTR regions. I've noticed, however, that the code assigning the intron number did not handle this case correctly and put the first intron as "0". This problem has been fixed for the OFCC1 genes on the public server. The fix will be public for all genes when we release the next build in a week or two.(**09/06/07**)

The Gene View section of the refSNP report for rs2461838 contains a table that has three "reference" contigs; the yellow rows that indicate the SNP is intronic, yet the pink rows that say it's coding nonsynonymous, resulting in a missense amino acid. I'm confused.

The RefSNP report for rs2461838 shows that the SNP maps to two different assemblies: the NCBI reference assembly and the Celera assembly. I think you may be confused by the second column, which shows the contig accessions; the mRNA accessions in the region of the contig where this SNP is found; and the protein accessions. If you look carefully at the first number in the second column, you'll see that there is just one

reference contig that this SNP maps to: NT_010718, but there are three different mRNAs in this region on the contig (the second number). The SNP happens to be intronic in the two NM_ accessioned mRNA products, while the XM_932588 mRNA is a predicted model and in this model, the SNP is missense (see the RefSeq FAQ or the RefSeq accessions page for the difference between NM_ and XM_ accessioned mRNA).

The following is a direct quote from the RefSeq FAQ page:

"...NCBI Accession numbers that begin with the prefix XM_ (mRNA), XR_ (non-coding transcript), and XP_ (protein) are model reference sequences produced by NCBI's Genome Annotation project. These records represent the transcripts and proteins that are annotated on the NCBI Contigs and they may be different from GenBank submissions for mRNAs and/or the curated RefSeq records (with NM,NR,NP accession prefixes). These differences may reflect real sequence variation (polymorphism), or errors (or gaps) in the available genomic sequence. These model RefSeq records should be used with caution, after comparing them to other available sequence information (Check the evidence viewer, BLink, Gene, or sequence neighbors)..."

You can get a key to the table colors by clicking on the words "color legend", located in the "Gene Model" column header, which is just below the blue "Gene View" section head.(**11/02/07**)

Why does the integrated maps section of the rs9288952 refSNP report show the alleles to be A/G, but the Geneview section shows them to be C/T?

The allele for rs9288952 was reported in refSNP (rs) flank orientation; this SNP also aligns to contig in the same orientation. The mRNA (NM_181780), however, aligns to the contig in reverse orientation. The allele reported in SNP functional class was reported using the mRNA orientation — this is the reason why the allele was listed as C/T. I have added a "hover title" to column header "dbSNP Allele" to make the orientation mentioned above more clear. The updated CGI will be public next Monday. (**4/20/07**)

The position information for non-synonymous SNPs in dbSNP's GeneView entry for NEU3 indicates at least 614 amino acid residues, but the reference protein sequence (NP_006647) shows only has 451 residues.

dbSNP annotated gene NEU3 onto NP_006647.2, but NP_006647.2 was replaced by a updated reference protein sequence (NP_006647.3) on Mar.25, 2007.

Go to Entrez Protein, and enter "NP_006647.2" (omit the quotation marks) in the text search box at the top of the page, and click the "Go" button. This will provide you with the record for NP_006647.2. Now use Entrez protein to find the records for NP_006647.3 using the same steps. When you compare the two data sets, you can see that NP_006647.3 is 461 amino acids long, while NP_006647.2 is 629 aa long. I will update the Gene view page and add the new version to all occurrences of Contig/protein/mRNA accessions. (**5/10/07**)

The refSNP (rs) reports for the clusters I'm interested in show that the variant positions match NP_055693.1, yet the Entrez Protein results and the XML files for these clusters both match NP_055693.4.

If you enter "NP_055693.1" into the text search box at the top of the Entrez Protein page, the resulting report shows that dbSNP b127 was indeed mapped in Dec, 2006 when NP_055693 was on version 1. A comparison of these two versions of NP_055693 shows that the protein sequence has changed significantly between version 1 (length 508) and version 4(length 648).

We are aware that a small percentage of protein and mRNA versions go out of sych between dbSNP and Entrez, and we are working on resolving this problem. If you subscribe to dbSNP-announce, you will be notified when this problem is solved. (**5/23/07**)

Clicking on SNPs located in the "Graphic display" for the reference mRNA model of rs2534719, 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)**

The amino acid assignments for rs1799971 (N102D) , rs1799972 (A68V), rs17174829 (D336N) are incorrect. You can see on NP_000905 that these amino acids actually occur at positions 40, 6, and 274.

All SNP protein positions are annotated for build 127 which came out on Mar, 8 2007. At that time, protein NP_000905 was version 2, having 462 amino acids. On July 29, 2007, however, the protein, NP_000905, was updated to version 3 and now has a length of 400 aa. This explains the positional difference of 62 amino acids.

You can see the protein and mRNA version numbers displayed on the GeneView page for the OPRM1 gene. Notice that it says "NM_000914.2" in the mRNA column, and "NP_000905.2 " In the protein column(the digit following the decimal is the version number); the version information is missing on the RefSNP page, a problem we will fix ASAP. We will also update the GeneView page so that it displays a line of text clearly stating the date that the SNPs were annotated to the mRNA and the protein so the user will at least know that some proteins may have changed after the SNP annotation was done.

The larger issue of keeping up with changing proteins, we will discuss within the SNP development group. On one hand, it is nice to update an annotation when the mRNA or protein changes, but on the other hand, many users like to use data from a specific SNP build. (**08/13/07**)

Integrated Maps Section

In the "Integrated Maps" section of rs55956772, it appears that each assembly maps the refSNP to different chromosomes without overlap. Is this due to small flanking sequence?

rs55956772 is from an Affymetrix chip and it was submitted with short flanks (16bp on each side). The shorter flank contributed to mapping difficulties. We have noticed such problems appearing in dbSNP and have begun to address them with a number of different approaches, one of which is the utilization of links to other supporting data from the submitter. (**07/14/08**)

rs6523677 is annotated as "A" in the reference sequences, but the genotyping (HapMap) of 300 individuals shows a "C" for all alleles. It's very unlikely that genotyping of any 300 individuals will show a "C" allele 300 times.

When dbSNP has genotyping data with overlapping samples from multiple submitters, we flag genotype conflicts. In this case, the only genotype data we have are from HapMap, so there is no additional validation available at the dbSNP site. If you look at the lower yellow box on the HapMap page for rs6523677, you will see the protocol used for the genotyping in question: urn:LSID:illumina.hapmap.org:Protocol:Golden_Gate_assay_design_1.0.0:1 (BeadArray platform).

We have noticed genotyping result differences between HapMap and direct sequencing in the past. Most of the differences are similar to the present one in that HapMap shows homozygous genotypes for a population while other sequencing methods turns up heterozygous genotypes. I have copied HapMap Help to see if they can comment or have any other information on this particular genotype. **(11/08/07)**

rs6523677 is annotated as an "A" in the reference sequences, but in the population genotyping provided by HapMap a "C" is found for all alleles. What's wrong with this dbSNP entry?

rs6523677 shows the genotype "C/C" for all populations on the HapMap site as well, which is expected since the geniotype data displayed for rs6523677 in dbSNP was submitted by HapMap. When dbSNP has genotyping data for a refSNP from multiple submitters with overlapping samples, we will flag genotype conflicts. In this case, the only genotype data we have are from HapMap. So there is no additional validation shown on the dbSNP site. You can find the protocol used by HapMap by scrolling down to the "Available Assays" section of the HapMap report for rs6523677.

We have noticed genotyping result differences between HapMap and direct sequencing in the past. Most of the differences are similar to the above case where HapMap shows homozygous results for a population while other sequencing methods show heterozygous genotypes.(**11/08/07**)

The record for rs7544137 states: "NCBI MapViewer: rs7544137 was not linked to the human genome 36.2 because it aligned to more than two locations on the genome", but there are only two contigs listed, and only two "CTG" records in the FLT file.

The statement: "rs7544137 was not linked to the human genome 36.2 because it aligned to more than two locations on the genome" is not accurate. Actually, rs7544137 was not linked to human genome 36.2 because it had a "map weight" value of three ("greater than two" in the statement in question).

A map weight of three indicates that a SNP does not uniquely map as it aligns at ≥ 10 locations, and as such is not linked to Mapviewer. There are map weight definitions available online. Once you're on the page, Scroll down a little to find them.

The above "not linked" statement is confusing, and will be updated to more accurately reflect reason the SNP was not linked to the genome. (**08/12/07**)

The refSNP report for rs749720 shows a chromosome base position (1:2379562) that is inconsistent with data for the same rs number returned by eutils (1:2379561).

The refSNP page shows sequence positions determined by counting the first base as 1, while internally, dbSNP counts the sequence position starting from 0 (the first base has position 0). In general, sequence positions are counted starting with 1 for user reports, while for programming exchange, sequence positions are counted starting with 0. Therefore, etuil returns data with positions counted from 0. To see the announcement indicating which counting system different file types use, go to the announcement section at the top of the dbSNP home page and scroll down to the bottom and click on the entry" 10/31/2005: 1 or 0 Based Mapping Position", and scroll down a little more to see details about the counting systems. (**5/11/07**)

Why does the integrated maps section of the rs9288952 refSNP report show the alleles to be A/G, but the Geneview section shows them to be C/T?

The allele for rs9288952 was reported in refSNP (rs) flank orientation; this SNP also aligns to contig in the same orientation. The mRNA (NM_181780), however, aligns to the contig in reverse orientation. The allele reported in SNP functional class was reported using the mRNA orientation — this is the reason why the allele was listed as C/T. I have added a "hover title" to column header "dbSNP Allele" to make the orientation mentioned above more clear. The updated CGI will be public next Monday. (**4/20/07**)

A certain refSNP Cluster Report states: "Integrated Maps: NCBI MapViewer: rs# was not linked to the human genome 36.2 mapviewer." What does this statement mean?

The statement means that the SNP was not linked to the genome since it does not have a unique map position to NCBI reference assembly 36.2.

This particular SNP has a mapping weight (a number that represents the mapping quality of the SNP on each assembly)of 3, which indicates that this SNP does not uniquely map as it aligns at ≥ 10 locations, and as such is not linked to the genome. There are map weight definitions available online. Scroll down a little to find them. (**08/21/07**)

Population Diversity Section (Frequency data)

Minor Allele Frequency (MAF) found in the population diversity section of the refSNP report is very high in SNPs from different sets of samples from the same ethnicity. Why?

Unfortunately, I think each SNP must be separately evaluated to determine why two sample populations from the same geographic region/ population group have different allele frequencies.

Below is a list of some the explanations for this occurrence that when applied to each individual case, should help you understand the reasons behind it:

- **Sequence Differences**: Look for neighbor SNPs and molecule type (DNA vs cDNA).
- **Sampling Differences**: Population grouping too broad. Look at the sample size and descriptions.
- **Strand Differences**: Look for opposite symmetrical allele frequencies.
- **Platform Chemistry**: Look at the method description.

(**12/04/07**)

rs6523677 is annotated as "A" in the reference sequences, but the genotyping (HapMap) of 300 individuals shows a "C" for all alleles. It's very unlikely that genotyping of any 300 individuals will show a "C" allele 300 times.

When dbSNP has genotyping data with overlapping samples from multiple submitters, we flag genotype conflicts. In this case, the only genotype data we have are from HapMap, so there is no additional validation available at the dbSNP site. If you look at the lower yellow box on the HapMap page for rs6523677, you will see the protocol used for the genotyping in question: urn:LSID:illumina.hapmap.org:Protocol:Golden_Gate_assay_design_1.0.0:1 (BeadArray platform).

We have noticed genotyping result differences between HapMap and direct sequencing in the past. Most of the differences are similar to the present one in that HapMap shows homozygous genotypes for a population while other sequencing methods turns up heterozygous genotypes. I have copied HapMap Help to see if they can comment or have any other information on this particular genotype. **(11/08/07)**

For rs1064733, the allele frequency for C is 1.0, but no frequency is given for the G allele. The genotype and allele frequency reports show that the G allele is present, but at frequency of zero. Please interpret these results.

The submitter records section of the rs1064733 refSNP page shows that rs1064733 was created when a submitter whose handle was LEE submitted ss1553687 in cDNA. The submitted SNP has an observed allele of C/G, but has no frequency data, and as you have pointed out, the subsequent frequency and genotyping studies (by HapMap) did not find a G allele. This SNP, therefore, is still unvalidated, which is represented in the submitter records section by a blank space below the "Validation Status" column header (You can click on the "Validation Status" column header to see dbSNP's validation criteria).

You can access the method used for discovering the SNP by clicking on submitted SNP (ss) number located at the far left column of the submitter records section. This will take you to the Submitted SNP detail page for ss1553687. Once there, scroll down and click on the words "Method –A" in the assay section of the page to see a description of the assay used to find the SNP You can see that this SNP was discovered using a computational method that "might produce spurious SNPs" according to the description.(**4/25/07**)

rs1801127 is validated by frequency data (according to the icon), yet has no heterozygosity data, while rs3219014 is not listed as validated but has frequency information.

1. Go to the refSNP page for rs3219014.

If you scroll down to the Validation Summary section at the bottom of the refSNP page, and click on the text "Validation Status" just below the section divider, you will see that the validation by frequency rule is "Validated by frequency or genotype data: minor alleles observed in at least two chromosomes."

If you scroll back up the refSNP page for rs3219014 to the "Submitter Records for this RefSNP Cluster" section and click on the only member (as of this date) submitted SNP (ss) (ss4480378) for this cluster, you will get the submitted SNP detail report. This report indicates that there is only one member of the population with the "A" allele (found in genotype "A/G") for all 90 people in the PDR90 population. The remaining members of the population have genotypes "G/G" or "N/N" (indeterminate).

As the Validation-by-frequency rule as shown in the first paragraph states that the minor allele count should be two or greater, rs3219014 (ss4480378) is not considered validated since even though it has available frequency data.

Even though rs3219014 lacks validation, this doesn't mean rs3219014 is not real. rs3219014 could be a rare SNP. If dbSNP gets future genotyping submissions from different submitters, it may be these submissions might show that rs3219014 is indeed rare.

2. As for rs1801127, the frequency/genotype data were submitted on member ss5586811 of the cluster. Looking at the data for ss5586811, you can see that this also seems to be a rare SNP, as it does not meet the validation by frequency rule that the minor allele count must be greater than or equal to two, and as such we will remove the "Validated by Frequency" flag from this SNP in a later update. Please note, however, that this SNP does meet the first validation rule, which states that a SNP can be " Validated by multiple, independent submissions to the refSNP cluster". (**5/10/07**)

The refSNP (rs) reports for the clusters I'm interested in show that the variant positions match NP_055693.1, yet the Entrez Protein results and the XML files for these clusters both match NP_055693.4.

If you enter "NP_055693.1" into the text search box at the top of the Entrez Protein page, the resulting report shows that dbSNP b127 was indeed mapped in Dec, 2006 when NP_055693 was on version 1. A comparison of these two versions of NP_055693 shows that the protein sequence has changed significantly between version 1 (length 508) and version 4(length 648).

We are aware that a small percentage of protein and mRNA versions go out of sych between dbSNP and Entrez, and we are working on resolving this problem. If you subscribe to dbSNP-announce, you will be notified when this problem is solved. (**5/23/07**)

Validation Summary Section

rs3181457 is validated, yet neither of the member ss of this cluster is validated. How trustworthy is the validation for this cluster?

The genotype results for rs3181457 in the refSNP cluster report as provided by HapMap, show that European and Sub-Saharan African are homozygous on T. The "G" allele count is one in each of the Han Chinese and the Japanese populations. This SNP has a minor allele count of two — just meeting the minimum criteria for this SNP to be called "validated by frequency".

The ss24813116 assay data was submitted by SEQUENOM with neither genotype nor frequency information, so ss24813116 isn't flagged as validated in the refSNP cluster report. The reason this refSNP cluster shows up as "validated by frequency of genotype data" is that HapMap submitted genotype data separately for the rs3181457 cluster. When the genotype data was submitted by HapMap, we linked it directly to the exemplar for this cluster, which is ss24813116. You can see this genotype data if you do the following:

1. Click in the link for ss24813116 located in the “NCBI Assay ID” column of the “Submitter Records” section of the rs3181457 refSNP cluster report
2. Once you get the “Submitted SNP Detail” report for ss24813116, scroll down to the “Submitted Individual Genotype” section of the report to see the data.

Another possible source of validation information is to look at the method description for ss24813116, but unfortunately, no detailed method description was submitted by SEQUENOM in this case. The submitter only submitted a method ID called "disc_method-1". You can get more information on SEQUENOM’s SNP discovery method, by contacting them using the contact information provided in the “Submitted SNP Detail” report for ss24813116. Just click on the “SEQUENOM” handle link. (**11/15/07**)