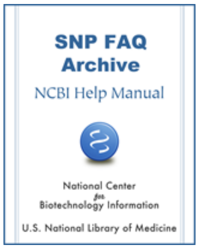
NIH U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: SNP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. Submitted SNPs (ss) and Other Data Submitted to dbSNP. 2005 Jul 23 [Updated 2014 Feb 18].
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/


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

Submitted SNPs (ss) and Other Data Submitted to dbSNP

Created: July 23, 2005; Updated: February 18, 2014.

Submitted SNPs (ss) Defined

What is the NCBI assay ID, or “ss” ID number?

The NCBI assay ID number or 'ss' ID number is simply a unique identifier in a standardized format (**NCBI|ss<NCBI ASSAY ID>**) that is assigned by NCBI to **s**ubmitted **S**NPs. Please note that “ss” is always in the lower case. (**04/05/06**)

Will a SNP keep its submitted SNP (ss) number once it has been assigned a refSNP (rs) number, or is the ss number no longer used?

A good explanation of the relationship between refSNPs (rs#) and Submitted SNPs (ss#) can be found in the "Computed Content" section of the dbSNP Handbook. Look in the "Submitted SNPs and Reference SNP Clusters" subsection. For more information regarding the assignment of rs numbers and citing unclustered ss numbers in a publication, look in the “Do I Need a RefSNP Number for Publication” sub section of the Submission section of the dbSNP FAQ archive. (**9/12/07**)

Why is there no SNP data for the factor VIII gene on the X chromosome? There are known polymorphisms in the gene.

Since dbSNP is a catalog of variations submitted mainly by our users, if a known polymorphism is not in dbSNP, it just means that Factor VIII has yet to be submitted. dbSNP adds new data daily, so perhaps SNPs for Factor VIII will submitted in the near future. Anyone can submit to dbSNP using our online instructions. (**9/25/07**)

Submitted Sequence Data

Helgadottir et al. indicates rs 10757278 and rs 2383207 are associated with coronary heart disease (CHD), while McPherson et al. indicates rs 10757274, and rs 2383206 are associated with CHD. Are the SNP pairs mentioned by the two groups the same?

These SNPs are not the same based on their unique flanking sequences and map positions on the genome. You can see this on Entrez SNP (**9/26/07**)

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

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

dbSNP maps subnmitted SNPs using the sequence of 5'_flank (+5'_assay) + observed + (3'_assay) + 3'_flank sequence. Inclusion of the assay sequences is optional, but if included, they cannot overlap with the flanking sequences. In your example, the flanking sequences for ss61470123 overlap the assay sequences. We will notify the user to correct the problem and resubmit the variation. (**6/8/07**)

I am trying to create primers for a list of rs numbers, and am looking for 50 bp of flanking sequence for these SNPs. Why does dbSNP show only 25bp of flanking sequence for these SNPs?

In general, dbSNP does not "chop" submitted sequence to only 25 bases of flanking sequence. 25 bases of flanking sequence are the minimum number dbSNP requires a submitter to provide when they submit a SNP assay to us. If a submitter only submitted 25bp of flanking sequence to dbSNP for the SNPs you are interested in, you could contact the submitter directly to see if they may be able to provide you with more flanking sequence.

If rs2000 was located 5 bases upstream of rs1000 in the genome, would each of these SNPs appear in the other's rs file even though they appear in each other's flanking sequence?

dbSNP does not currently report variations in flanking sequence. We'll look into providing this information in the future. **(10/19/05)**

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**)

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**)

Determining Orientation of Illumina Array Data

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

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

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

1. You can compute the top/bottom designation yourself using the data in the /organisms/human_9606/GWAS_arrays/ directory on the dbSNP FTP site.
2. You can look at dbSNP's top/bottom assignment, which you can access if you download the SubSNP.bcp file located in the /database/organism_data/ directory for human. The field that includes

the top/bottom data is called SubSNP.top_or_bot_strand. You can access the table DDL for SubSNP in the /database/organism_schema directory.
The downside of this approach is that you need to download the entire SubSNP table, which includes 50million+ submitted SNPs.

(**09/22/08**)

Species Data included in dbSNP

What kind of plant species will dbSNP house in the future?

It is possible that dbSNP will eventually house all plant species that have a genome project. There is an online list of plant genome mapping projects you can review. In the immediate future, the bulk of the SNP submissions will likely be from major crop plants (e.g. rice, corn, soy bean, and wheat). **(3/11/05)**

Do you have information about the breed of dog used as the source of the SNP dog data?

All our *Canis familiaris* SNPs are from a TIGR poodle. Searching by "tax_id 9615" only yields one batch from TIGR.

A description of the SNP population derived from a poodle named "Shadow" used in the TIGR submission is available.

Ethnic Data in dbSNP

How do I find out what ethnic groups were sampled for human SNPs housed in dbSNP?

Most of the genotype data in dbSNP are from HapMap, and their samples are from four populations. Information about these populations are available at the International HapMap Project Site.(**01/03/08**)

Strain Data In dbSNP

What were the rat strains used in the SNP comparisons for ss16343960 through ss16355610?

On the basis of the method the submitter provided to dbSNP with their submission,
it looks like they compared each strain against the rat draft genome sequence (Rnor3.1).

Submitter Information

rs28935498 has little (or no) detail regarding the submitter (OMIMSNP), population frequency, references, validation type. Do you have any additional information?

rs28935498 was mined programmatically from OMIM records and sequence alignments. The local SNP ID: OMIM_305400_0005 indicates that it is from OMIM record 305400, and its variant ID is 0005. Please note that the link on OMIM record 305400 shows that this record has been assigned a new OMIM ID: 300546. Here is the text description for variant 0005 from the OMIM report for 300456:

.0005 MENTAL RETARDATION, X-LINKED NONSYNDROMIC [FGD1, PRO312LEU]

Lebel et al. (2002) described 3 brothers with nonsyndromal X-linked mental retardation and a pro312-to-leu (P312L) missense mutation in the FGD1 gene. Although the brothers had short stature and small feet, they lacked distinct craniofacial, skeletal, or genital findings suggestive of Aarskog syndrome. Their mother, the only obligate carrier available for testing, had the FGD1 mutation. A 934C-T base change in exon 4 was responsible for the P312L amino acid substitution. This missense mutation was predicted to eliminate a beta-turn, creating an extra-long stretch of coiled sequence that may affect the orientations of the SH3 binding domain and the first structural conserved region.

That's all the information we have on this variation; you may wish to read the primary publication (Lebel et al.).(**01/11/08**)

Submitter Documentation for SNPs

I have found three different types of submittedIndIds: SubmitInfo popId="837", submittedIndId="SPRETUS", and submittedIndId="Western Wild Mouse" Are these all examples of mouse strain names?

Yes, each is a strain name used by the submitter to report a strain genotype. **(6/2/05)**

Is there a relationship between the numeric version of loc_ind_id and the characterized version?

loc_ind_id is the identification that a submitter gives to an individual submitted unit, whether it a person, sample, organism, or cell line. That's why it is a character field. dbSNP has an internal ind_id that acts as a unique integer identifier for a person.

Whenever possible, we map the ind_id to a Coriell number (for humans) or to a Jax Lab number (for mouse). Internally, we call this map the "source_ind_id" or "source". When we are unable to map a submitted individual to a source (Coriell or Jax for now), we use the Handle and loc_ind_id.

HapMap Data

Where do I find SNP data from the second phase of the HapMap Project recently published by the HapMap consortium?

The phase II HapMap genotype data (see the PubMed abstract of the phase II data publication in Nature) is currently available in dbSNP.

This genotype data is from HapMap release 21a which contains both phase I and phase II data. There are a number of ways to access this data from dbSNP: you can use Entrez SNP, eUtils, or an FTP download depending on whether you need the entire data set or data specific to particular genes or chromosome regions.(**11/27/07**)

How do I get the age of the 270 HapMap samples?

dbSNP does not have age information for the HapMap samples. What we do have is the pedigree information for those individuals. Since all 270 samples for the HapMap project have Coriell sample IDs, you could also view information for these samples on the Coriell site. Here is the Coriell Data summary for Coriell ID#: HAPMAPPT07

(**10/23/06**)

rs13406935 has two submitted SNPs (ss), and some HapMap data referenced to one of the submitted SNPs. Why doesn't the HapMap data for this refSNP have its own unique ss number?

The HapMap project used refSNP(rs) space for their project, and didn't actually submit genotypes back to the dbSNP in the traditional pipeline — instead, we uploaded their genotype release output directly to dbSNP. Once we loaded their genotypes we assigned the exemplar ss number as the HapMap assay ID. Each HapMap assayID was then loaded to the Entrez Probe database where we linked each genotype to a probe.

Assigning the exemplar ss as the HapMap assay ID and loading these Assay IDs into the the Entrez Probe database will be our new paradigm for genotype submissions that submit data for existing refSNPs(rs) as opposed to plain submitted SNP (ss) data. The Entrez Probe group will be independently mapping the probes to subsequent genome assemblies. dbSNP will sync with the probe group to compare mapping of the probes and SNPs for agreement. When there is disagreement, dbSNP will provide this information to the genotype submitters — the HapMap analysis group in this case. To count the number of rs numbers typed by HapMap, join the SubInd, Batch and SNPSubSNPLink tables where handle = "CSHL-HAPMAP"(**6/19/06**)

Computationally Derived SNPs

rs28935498 has little detail regarding the submitter (OMIMSNP), population frequency, references, validation type. Do you have any additional information?

rs28935498 was mined programmatically from OMIM records and sequence alignments. The local SNP ID: OMIM_305400_0005 indicates that it is from OMIM record 305400, and its variant ID is 0005. Please note that the link on OMIM record 305400 shows that this record has been assigned a new OMIM ID: 300546. Here is the text description for variant 0005 from the OMIM report for 300456:

.0005 MENTAL RETARDATION, X-LINKED NONSYNDROMIC [FGD1, PRO312LEU]

Lebel et al. (2002) described 3 brothers with nonsyndromal X-linked mental retardation and a pro312-to-leu (P312L) missense mutation in the FGD1 gene. Although the brothers had short stature and small feet, they lacked distinct craniofacial, skeletal, or genital findings suggestive of Aarskog syndrome. Their mother, the only obligate carrier available for testing, had the FGD1 mutation. A 934C-T base change in exon 4 was responsible for the P312L amino acid substitution. This missense mutation was predicted to eliminate a beta-turn, creating an extra-long stretch of coiled sequence that may affect the orientations of the SH3 binding domain and the first structural conserved region.

That's all the information we have on this variation; you may wish to read the primary publication (Lebel et al.).(**01/11/08**)

Although rs28935498 was mined from OMIM record 300546, and the reference mentions a mutation (P312L), the record does not indicate that this mutation is a SNP, and I can find no submissions of P312L as a SNP.

dbSNP was originally intended to store Single Nucleotide Polymorphisms (SNP), hence the name dbSNP. dbSNP, however, has evolved into a general variation database storing ANY genetic variation. Some of these variations come from mutation databases and we maintain the connection to the original records as evidence. Please see the online FAQ for uses of the term "SNP" and classes of genetic variation in dbSNP. (**01/22/08**)

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**)

Published Mutations in dbSNP?

We are surprised to find a SNP (rs28928906) for the MPI gene (NM-002435), since the non synonymous substitution is a published mutation, and a homozygous R295H patient had a deficient PMI activity.

This SNP was included in dbSNP as part of our on going effort to integrate variations from the OMIM database into dbSNP, so as to create connections between the variation textual information and references. The OMIM records for this SNP are available online.

If you wish to contribute your annotation for this variation to dbSNP, you can do so by using the Variation Batch Submission page. If you would like information about using this online submission tool, please see the VBS Quick Start. **(05/06: 12//07: 07/25/08)**

Validation Data

Definition of Validation Types

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**)

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 a 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)**

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 criteria you have for your four validation parameters—cluster, frequency, submitter, and double hit?

You can access this information in the SNP handbook. The handbook, however, doesn't contain double hit information, but you can find it in the dbSNP docsum spec, which states: "doublehit (3)—refSNPs with both alleles seen twice. Data for NSE-rs.validated-by-2hit-2allele."

Does dbSNP check that submissions in the multiple reporting validation method are independent of each other?

"By Cluster" validation is currently the only check used that will confirm that an rs cluster has at least two subSNPs (ss), and that at least one of the ss is derived from a non-computational method. Although we could check to determine if the ss are from two different submitters, it may not be possible because some submitters do not report this information.

How do you decide what validation status to use for a SNP?

You can access this information in the SNP chapter of the NCBI handbook.

How can validation status go from validated to not validated?

A change in validation status can happen in the following cases:

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 a 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 associated with lost its validation status.

When there is a bug in dbSNP that accidentally makes a rs "validated", and this bug is later fixed, the rs status changes back to "not validated. **(2/11/05)**

Interpreting Validation Discrepancies

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" (indetermindate).

   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**)

dbSNP shows a SNP as having been "validated by the HapMap project", but the genotype data shows this SNP to have just one allele in all 269 samples. How will dbSNP represent such conflicting genotyping results, and how will this play into the notion of refSNP "validation status"?

We are considering a new processing rule that does not allow monomorphic refSNPs to be validated. A HapMap flag on the validation really indicates that the SNP was "genotyped by HapMap". Our current validation rule deliberately makes HapMap "validation" and "validated by having minimum of two minor alleles", two separate validation conditions. It is easy to see how those refSNPs with HapMap genotype data would turn out to have no variation when you use a simple query such as: Select count(*) from SNP where validation_status-16 = 0 62479
(4/20/05)

```
Examples are:
snp_id
-----------
26441
26716
26733
26943
27073
27490
29151
30970
31411
31742
```

Extending Validation to Protein Level

Can I assume that if a SNP is validated, it is also validated in all the proteins in to which it maps?

SNPs are experimentally validated through observation, and because SNPs are observed as variations in the nucleotide sequence, SNPs are experimentally validated at the nucleotide sequence level. The amino acid variations are predicted from their corresponding nucleotide variations — they have not been confirmed by protein sequencing — therefore they cannot be considered experimentally validated. **(6/14/05)**

Validation of SNPs in Specific Builds

How were SNPs in Build 121 validated?

You can find this out for yourself. As an example of how to accomplish this, look in the second section of the page, called Submitter Records by RefSNP Cluster. Click on the words Validation Status (the heading of the third field of that section), and you will see a pop-up window containing the validation codes. You can also see the validation status code definitions on the dbSNP FTP site.

Purging Non-Validated SNPs

I have noticed many SNPs that are not validated. Why doesn't dbSNP purge these SNPs?

Currently, dbSNP does not unilaterally remove SNPs that are not validated. We will, however, process a submitters' withdrawal request if they later find their submission to be an artifact.

dbSNP also does not remove “un-validated” SNPs since it is possible that a SNP derived computationally may later be successfully genotyped by another lab.

If you are interested in using an SNP for your research, you should first make an informed decision about the usefulness of the SNP by reviewing the validation status (found on the refSNP cluster report) and detection method (links for available detection methods can be found on the submitted SNP detail report) of the SNPs under consideration.

(**12/06/07**)

Population Diversity Data

Allele Data

How does dbSNP code alleles for SNPs from multiple studies across multiple platforms so that they are comparable?

A submitted SNP has a fixed orientation that is defined by its flanking sequence which allows us to map the alleles to the refSNP, the reference genome, or any number of genomic sequences/assemblies.

Top / Bottom orientation can be determined by a simple algorithm applied to the alleles and flanking sequence, however, the algorithm is defined for single base alleles only and may not work with some palendromic flanking sequences. *Note*: To orient linked SNPs in a haplotype, you cannot use submitted SNP (ss), refSNP (rs) or top/bottom orientations; instead you must use the orientation to a common sequence or assembly. (**08/08/08**)

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**)

Genotype Data

The genotype results for rs34958084 shows are strange: in the ss48428804 assay, 100% of CEU, HCB and JPT individuals are homozygous for T; but, in the assay ss66405533 assay, 100% of CEU, HCB and JPT individuals are homozygous for C.

You will see as you examine the reports for rs34958084, that the submitter of the genotype data was "GAIN-BROAD-QC".

We have noticed the genotype data quality issues for a number of batches from "GAIN-PERLEGEN-QC" and "GAIN-BROAD-QC", and one of our staff members has been working with these submitters to fix the genotype data. We hope the data will be corrected in time to be released with the next build (B130). (**07/17/08**)

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**)

How do I verify dbSNP's genotype data sources?

Population descriptions are provided to dbSNP by the submitter, and as such, the submitter is entirely responsible for any quality control of the submitted population description. You will therefore have to ask the submitter about the source of the genotype data. Please contact snp-admin@ncbi.nlm.nih.gov if you need help contacting the submitter.

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 — 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)**

HapMap Data

For rs28376053 and other refSNPs where two or more datasets were submitted for the same SNP in the same HapMap samples, genotypes were "reversed" in one dataset.

There are individual genotype conflicts are between two submitters (GAIN and HapMap). The GAIN project has submitted sets of genotypes by two different centers(BROAD and PERLEGEN) for a subset of HapMap genotyped SNPs using the same HapMap samples as a Quality Check process. We have since found conflicting genotype results (for the same SNP in the same individual).

A member of the IEB Genome Variation Working Group here at NCBI has analyzed the QC data from GAIN and gave us his results: For rs28376053, GAIN submitted genotype data for ss66115729 that was confirmed as correct. This suggests that the HapMap release 23 genotype results for this SNP are incorrect. I will implement a "genotype conflict warning" function which will flag such conflicts in a future build.

As we are working to load HapMap phase 3 data, I checked to see if phase 3 produced different genotype data for **rs28376053,** but phase 3 evidently did not include rs28376053. (Phase 3 typed 1.6 million SNPs over 11 populations, while HapMap phase 2 [such as release 23] typed 4 million SNPs over 4 populations). So until we get new genotype data from HapMap on rs28376053, this genotype conflict will stay in dbSNP, but will eventually have a conflict warning flag. (**08/13/08**)

I've noticed that in a number of SNPs (e.g. rs34950166 and rs35040247) all individuals examined in the 4 HapMap populations are always heterozygous. This seems unlikely.

The refSNPs you mentioned were genotyped in a QA project by PERLEGEN using HapMap samples. For example, if you go to the genotype section of the refSNP cluster report for rs34950166, and click on ss68759579, you will see in the page below that the genotype data was submitted whose submitter ID (handle) is "GAIN-PERLEGEN-QC".

We have noticed the genotype data quality issues for a number of batches from "GAIN-PERLEGEN-QC" and "GAIN-BROAD-QC", and one of our staff members has been working with these submitters to fix the genotype data. We hope the data will be corrected in time to be released with the next build (B130). (**07/17/08**)

Where do I find SNP data from the second phase of the HapMap Project recently published by the HapMap consortium?

The phase II HapMap genotype data (see the PubMed abstract of the phase II data publication in Nature) is currently available in dbSNP.

This genotype data is from HapMap release 21a which contains both phase I and phase II data. There are a number of ways to access this data from dbSNP: you can use Entrez SNP, eUtils, or an FTP download depending on whether you need the entire data set or data specific to particular genes or chromosome regions.(**11/27/07**)

Frequency Data

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**)

Does dbSNP use the summary of genotypes table to produce the submitted frequency table, or vice versa?

dbSNP summarizes the submitted genotypes and shows the results in the table dbSNP summary of genotypes, whereas the numbers in the submitted frequency table are provided by the submitter.

Why are four nucleotides listed for the average allele frequency of rs702424?

I checked the submission records and found inconsistencies in the original data. rs702424 has two ss with frequency data: ss1724861 and ss152049; both of these are located on the forward strand of the rs. The ss152049 frequency data is for "T" and "C", while the ss1724861 frequency submission was on the wrong strand. The submitter listed "A" and "G" on the FORWARD strand of ss1724861in the frequency file, but listed "A" and "G" on the REVERSE strand of ss1724861 in the genotype file. The frequency for ss1724861 should be listed as "C" and "T" on FORWARD strand, or "A" and "G" on the REVERSE strand. To reduce the effect of submission error in our computation of allele frequency, we have recently started to use the following rules:

For the same ss number and population:

- When there is genotype data, we ignore the genotype frequency and allele frequency data.
- When there is only genotype frequency data, we ignore allele frequency data.
- When there are different submitted allele frequencies, we choose the one which has the largest sample size.
- When different frequency data are submitted with same sample size, we ignore them.

The results computed using the above rules are kept in a table named SNPAlleleFreq, which is located in the organism_data directory (this link is to the human_9606 directory) of your organism's database. The data on the wrong strand ("A" and "G") was considered in calculating the average allele frequency since the web page was still using all data sources at the time the calculation was performed.

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**)

Frequency Data Discrepancies

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" (indetermindate).

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 first entry in the submitted frequency table for ss870160 shows 100% of the 184 CEPH chromosomes have a "T" allele, but the second entry shows 100% of 84 TSC_42_C chromosomes have an "A" allele. Why?

It is most likely that KWOK submitted their frequency data on the opposite strand of the one used to discover the SNP. In the early days of dbSNP, we accepted frequency data without strand information.

We have since required that all allele frequency or genotype submissions must be accompanied by strand information (e.g. SS_STRAND_REV, SS_STRAND_FWD). We are currently requesting data submitters that do not have strand information to either give us individual genotype data to replace the allele frequency, or to provide the strand information.

In the meantime, we will "flag" possible strand errors in our current allele frequency data. **(8/5/05)**

Sample Ascertainment

Population (2N) Sample Size

The "Sample (2N)" or Population Sample Size for rs1800849 is 1767. Why is this an odd number? If it is based on the number of chromosomes, the number should be even.

The 2N sample size, or population sample size is reported by the submitter, along with his/her population frequency data, and is the number of chromosomes that the submitter used as the denominator in computing estimates of allele frequencies.

The odd number for the population sample size indicates that some of the chromosomes must have failed the experiment. To be sure, I have forwarded your question to the submitter and will forward the answer you when we hear from them.

To see the data for the original submission, see the ss page. Go to the Resource Links section, located at the top right of the page. At the bottom of the Resource Links section is the submission report field. Click on the word view to see the submission report.

For rs660339, why is the population sample size (2N) for population SC_12_C in the summary of genotypes table different from that in the submitted frequency table?

When we have individual genotype and pedigree information, we exclude non-founders (individuals whose mother or father were also sampled for the same SNP) in calculating genotype frequency (displayed in table dbSNP summary of genotypes).

Population SC_12_C was sampled for this SNP and has a total individual count of 12, but 5 of these individuals are non-founders:

Non founder for this ss/pop: CEPH1416.03 CEPH1416.02 CEPH1416.01
Non founder for this ss/pop: CEPH1416.04 CEPH1416.02 CEPH1416.01
Non founder for this ss/pop: CEPH1347.03 CEPH1347.02 CEPH1347.01
Non founder for this ss/pop: CEPH1347.04 CEPH1347.02 CEPH1347.01
Non founder for this ss/pop: CEPH1347.05 CEPH1347.02 CEPH1347.01

Please note that the population sample size (2N) in the dbSNP summary of genotypes table and in the submitted frequency table for populations SC_12_A and SC_12_AA agree because all of the individuals in the samples were founders.

Total Sample Size

How do I find the total number of samples (i.e. people) you’ve sequenced for a particular polymorphism of human maspin (serpin b5)?

dbSNP does not generate the SNP data. dbSNP is a depository for data submitted from hundreds of research groups. Each SNP may have a different sample size from another SNP. Search for rs2289519 using Entrez SNP. The results page for this search shows a row of colored buttons below rs2289519 that represent links. Click the pink "GeneView” button. This will take you to the “SNP linked to Gene” page, which shows that P176S corresponds to refSNP rs2289519. The total sample size for rs2289519 is located in the Population Diversity Section of the refSNP report.(**6/1/06**)

Population Source Data

Submitter details for some genotype data that I was examining indicate that the genotypes were from North Americans. Is it certain that the population tested was North American only?

Population descriptions are provided to dbSNP by submitters, and as such, the submitter is entirely responsible for any quality control of the submitted population description. You will, therefore, have to ask the submitter about the source of the genotype data. Please contact snp-admin@ncbi.nlm.nih.gov if you need help contacting the submitter.

What was the population used to determine the frequency for rs6267? Since this SNP tends to occur more often in Asians than in Caucasians, I would like to know.

Below is a list of all frequencies by populations. The list has 6 fields: handle, submitter_pop_id, pop_id, a.allele, s.cnt, s.freq. This rs number has frequency data from six different submitters.

–

Handle	submitter_pop_id	pop_id	a.allele	s.cnt	s.freq
NCBI	NIHPDR	506	G	71.000000	0.934211
NCBI	NIHPDR	506	T	5.000000	0.065789
SNP500CANCER	P1	754	G	196.949997	0.975000
SNP500CANCER	P1	754	T	5.050000	0.025000
SNP500CANCER	CAUC1	775	G	59.010002	0.983500
SNP500CANCER	CAUC1	775	T	0.990000	0.016500

_ *continued from previous page.*

Handle	submitter_pop_id	pop_id	a.allele	s.cnt	s.freq
SNP500CANCER	AFR1	776	G	46.992001	0.979000
SNP500CANCER	AFR1	776	T	1.008000	0.021000
SNP500CANCER	HISP1	777	G	45.010998	0.978500
SNP500CANCER	HISP1	777	T	0.989000	0.021500
SNP500CANCER	PAC1	778	G	46.007999	0.958500
SNP500CANCER	PAC1	778	T	1.992000	0.041500
EGP_SNPS	PDR90	693	G	154.000000	0.939024
EGP_SNPS	PDR90	693	T	10.000000	0.060976
SEQUENOM	CEPH	1303	G	184.000000	1.000000
SEQUENOM	CEPH	1303	T	0.000000	0.000000
CSHL-HAPMAP	HapMap-CEU	1409	G	120.000000	1.000000
CSHL-HAPMAP	HapMap-HCB	1410	G	83.000000	0.943000
CSHL-HAPMAP	HapMap-HCB	1410	T	5.000000	0.057000
CSHL-HAPMAP	HapMap-JPT	1411	G	82.000000	0.932000
CSHL-HAPMAP	HapMap-JPT	1411	T	6.000000	0.068000
CSHL-HAPMAP	HapMap-YRI	1412	G	120.000000	1.000000
PERLEGEN	AFD_EUR_PANEL	1371	G	48.000000	1.000000
PERLEGEN	AFD_AFR_PANEL	1372	G	46.000000	1.000000
PERLEGEN	AFD_CHN_PANEL	1373	G	43.000000	0.934783
PERLEGEN	AFD_CHN_PANEL	1373	T	3.000000	0.065217

You can view descriptions of submitted populations. To view a description of a particular population, substitute the pop_id number you are interested in at the end of the URL displayed for the above link. **(9/20/05)**

How do I verify dbSNP's genotype data sources?

Population descriptions are provided to dbSNP by the submitter, and as such, the submitter is entirely responsible for any quality control of the submitted population description. You will therefore have to ask the submitter about the source of the genotype data. Please contact snp-admin@ncbi.nlm.nih.gov if you need help contacting the submitter.