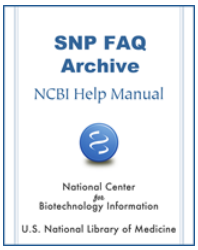
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: SNP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. Clustered RefSNPs (rs) and Other Data Computed in House. 2005 Jul 23 [Updated 2014 Feb 18].
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Clustered RefSNPs (rs) and Other Data Computed in House

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

RefSNPs (rs) Defined

What is a reference SNP, or "rs" ID number?

A **r**eference **S**NP ID number, or "rs" ID, is an identification tag assigned by NCBI to a group (or cluster) of SNPs that map to an identical location. The rs ID number, or rs tag, is assigned after submission. When dbSNP was first released to the public in 1998, every submission that appeared to be unique in the database was assigned separate rs ID numbers. Now that dbSNP has matured with constant submissions, a submitted SNP is evaluated to see if it maps to an identical location as previously submitted SNPs; if it does, then the submitted SNP is linked into the reference set of the existing reference SNP record. These SNP rs IDs are mapped to external resources or databases, including NCBI databases. The SNP rs ID number is noted on the records of these external resources and databases in order to point users back to the original dbSNP records. A reference SNP record has the format **NCBI| rs<NCBI SNP ID>**. Please note that 'rs' is always in the lower case. For further information on refSNPs, please see our online documentation.(**04/05/06**)

Is a "cluster" (as in "RefSNP cluster") a collection of all submissions of a particular sequence?

Yes. If multiple submitted SNPs (called "ss" for short; each is assigned a unique ss number) map to the same contig, we cluster these ss numbers into a refSNP cluster (called "rs" for short).

Does refSNP number rs1801280 represent the location of the SNP or the SNP itself?

The refSNP (rs) number is a unique identifier for SNP and does not infer position. An rs number may have one or more positions on a sequence (ie. SNP located in repeated region). When referring to a genotype for an individual, it should be specified with the rs number as well as the observed alleles (ie. A/A, A/T, or T/T). Please refer to dbSNP SNPINDUSE submission documentation if you need the format for submitting genotypes to dbSNP. (**06/05/08**)

Where can I find definitions for terms like "refSNP" and "validated SNP"?

You can find these definitions in the dbSNP chapter of the NCBI handbook. A link to the handbook is located on the blue left side bar of the dbSNP home page. (**4/21/06**)

Both rs6419492 and rs4601571 seem to describe the same SNP, yet both rs numbers exist separately in dbSNP. Am I missing something?

Yes, these two refSNPs should probably be merged. In general, we cluster submissions that co-locate on the genome, but our heuristics sometimes exclude particular cases. We prefer to err on the side of caution; better to leave a few co-located SNPs unclustered than cluster submissions which do not belong together. In this

case, I can see from the submitter comments that the two refSNPs are from the same genomic location. Thanks for bringing this to our attention. **(3/13/05)**

rs28937569 and rs28937568 are listed as disease-related mutations in OMIM. Can a mutation be listed as a variation in dbSNP?

Please Note: the use of the word "mutation" is being phased out in dbSNP, and will be replaced by the term "Clinical/LSDB variation".

rs28937569 and rs28937568 were submitted by OMIMSNP using an automated program that extracted OMIM variations occurring in the coding regions of OMIM records. Only those variations with reference sequences available and where the reference allele was confirmed were added to dbSNP.

Originally, the great majority of data in dbSNP was collected and defined as variations simply using sets of co-aligned genomic or DNA sequences. Because this process typically had little to no focus on disease condition, only about 250 records in dbSNP were successfully associated with phenotype-causing variations or a clinical outcome in OMIM.

Starting in the Spring of 2008, however, dbSNP began accepting submissions of Clinical/LSDB variations as well as annotations to existing variations (including phenotype) on the Human Variation: Search, Annotate, Submit site (for single submissions) as well as on the Human Variation: Annotate and Submit Batch Data site (for multiple submissions). As of this date, there are a total of 1266 records in dbSNP that were submitted as Clinical/LSDB variations (select "Clinical/LSDB variation" in the Entrez SNP limits page and click "GO" with out entering a search term in the Entrez search box), and 1134 records submitted as Clinical/LSDB variations that also have OMIM links (select "Clinical/LSDB variation" and "OMIM" in the Entrez SNP limits page and click "GO" with out entering a search term in the Entrez search box).

Those SNPs with clinical association(s) will have a red "VarView" (Variation Viewer) link in the "allele" section at the upper right of the refSNP cluster report. Clicking the link will take you to the Variation Viewer Report for the gene in which the SNP is found.

We expect that the number of Clinical/LSDB variation records in dbSNP will grow rapidly as more users discover dbSNP's resources for submitting them. (**07/09/08**)

RefSNP (rs) Characteristics

Do rs numbers have a maximum length?

Currently, there is no limit on the length of these integers.

Are refSNP (rs) numbers unique for each SNP and do they apply to all types of variants?

RefSNP (rs) numbers are unique and are used for all types of variants.

Hardy Weinberg Equilibrium Data

I am having trouble understanding how the Hardy-Weinburg Probability (HWP) value is calculated for SNP records. I'm no statistician, so it would be helpful if you could include an example.

A good explanation of HWP and how it is calculated can be found in an online Wikipedia article.

Currently, we use the chi-square and p-value in our HWP calculations. Here is an example of how we calculate the HWP for ss221 in a population (population id as 506) following the HWP equation.

–

GtyFreq	subsnp_id	pop_id	gty_str	cnt
0.411	221	506	A/G	37.000000
0.144	221	506	A/A	13.000000
0.444	221	506	G/G	40.000000

–

cnt	AlleleFreq	subsnp_id	pop_id	allele
117.000000	0.650	221	506	G
63.000000	0.350	221	506	A

–

subsnp_id	pop_id	ind_cnt
221	506	90.000000

–

subsnp_id	pop_id	chr_cnt
221	506	180.000000

–

subsnp_id	pop_id	Degree of Freedom
221	506	1

–

exp_gtyFreq	subsnp_id	pop_id	exp_gtyCnt	allele_1	allele_2	gty_str
0.422	221	506	38.025	G	G	G/G
0.455	221	506	40.950	G	A	A/G
0.122	221	506	11.025	A	A	A/A

–

subsnp_id	pop_id	gty_str	observed_gtyCnt
221	506	G/G	40.00
221	506	A/G	37.00
221	506	A/A	13.00

–

subsnp_id	pop_id	Chi-square
221	506	0.837

After obtaining the chi-square value, calculate the p-value using the degrees of freedom via the gamma function, or many websites offer p-value calculators. (**9/11/07**)

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 next question in this section also contains information you should find useful. (**6/4/07**)

Your Hardy Weinberg probabilities are calculated using 2 df—an incorrect calculation. Because only one independent parameter is being used to estimate (p=1-q), = there should only be 1 df.

Thanks for pointing out the inclusion of failed assays in the Hardy Weinberg estimates for refSNP (rs) clusters appearing on rs cluster reports as well as in the genotype and allele frequency reports. We will exclude such genotypes in the Hardy Weinberg equilibrium calculations for the release of dbSNP build 120.

Many of the assumptions of the Hardy Weinberg equilibrium are not necessarily met in the submissions to the dbSNP, but we do believe that there is utility (although limited) in calculation of Hardy Weinberg equilibrium over rs clusters. Please see the Hardy Weinberg equilibrium performed over submitted SNPs (ss) by population. These are listed in the submitted SNP details page as well as in the genotype and allele frequency report. Please be aware, however, that the population in the context of dbSNP may (or may not) differ from a population as defined in a population genetics context.

The displayed value for HWP on your web page is different than the FAQ which calculates it. Am I using the correct calculation method?

To speed up database update, we have used a lookup table which is binned. This is a contributory factor as to why the P value is not exact but should be close. To calculate the P value (chi-square distribution), we used the Gamma Function. I think we got the C algorithm from Numeric Recipe book.

If you are interested in HWE, have you looked at Fisher Exact test? From what biologists have told us, the Fisher Exact test is a better way to estimate HWP especially when sample size is small. When time allows, we hope to switch our HWE calculations from the Chi-square test to the Fisher Exact test.(**08/01/08**)

Merging RefSNP Numbers and RefSNP Clusters

I heard that RS numbers are not stable. For example, rs17216163 is now rs717620, and rs17231380 is now rs5186. I assume you don't ever reuse the "retired" numbers? Why did you make some numbers obsolete?

The examples you cite are instances where multiple rs numbers were assigned at the same genomic location, and the higher rs number was merged into a lower rs number (this is the dbSNP merge rule for rs numbers). Such a merge can happen when submissions differ in the length and quality of flanking sequence. We only merge rs numbers that have an identical set of mappings to the genome and have the same type of alleles (e.g. both must be the same variation type and share one allele in common). We would not merge a SNP and an indel (insertion/deletion) into a single rs number (different variation classes) since they represent to different types of mutational "events".

The location of the rs number remains valid and we never reuse rs numbers.

We have discussed the issue of supporting query by merged rs numbers more robustly in dbSNP, Entrez and our web based services. That way a retired rs number can be found easily and used as a proxy for the current "live" number. Please note that merging is only used to reduce redundancy in the catalog of rs numbers so each position has a unique identifier.

In the first example you cite, prior to their merge, both rs17216163 and rs717620 would have been the "address" for the same nucleotide. Now only rs717620 is used in annotation, and rs17216163 is retained in our merge history tables. With extended annotation, users would be able to query by the full set of retired rs numbers.

Currently, there are three different entry points in dbSNP that will lead you to the partner numbers of a merge:

1. You can retrieve a list of merged rs numbers from Entrez SNP. Just type “mergedrs” (without the quotation marks) in the text box at the top of the page and click the “go” button. Each entry in the returned list will include the old rs numbers that has merged, and the new rs number it has merged

into (with a link to the refSNPpage for the new rs number). You can limit the output to merged rs numbers within a certain species by clicking on the "Limits" tab and then selecting the organism you wish from the organism selection box.

2. If you enter a merged old rs number into the "Search for IDs" search on the dbSNP home page, the response page will state that the SNP has been merged, and will provide the new rs number and a link to the refSNP page for that new rs number.
3. The RsMergeArch table houses the merged SNPs, and is available on the dbSNP ftp site. A full description of the table can be found in the dbSNP Data Dictionary, and the column definitions are located in the dbSNP_main_table.sql.gz, which can be found in the shared_schema directory of the dbSNP FTP site.

(**11/07:05/08:11/08**)

Why is rs8111802 classified as a SNP rather than "mixed" when its exemplar submission is a DIP record?

The SNP development group thinks that it is best if we do not cluster or merge SNPs of different variation classes even when they map to the exact same contig location. This affects about 50K of the current refSNP (rs) numbers, including rs8111802. We will split these current SNP clusters by SNP class and work out all related details soon.

(**10/6/06**)

dbSNP has more tetra-allelic SNPs than one would expect. For example, rs1045642 has 4 alleles (a/c/g/t), but in each population only two show up (C/T or A/G).

You are correct — most of the tetra-allelic SNPs in dbSNP are the result of cluster orientation error. There are a total of 2361 human SNPs that are tetra-allelic. We are working on re-blasting these SNPs and correcting this problem. **(9/19/05)**

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)**

Can two ID numbers correspond to the same SNP?

If you mean "can two SNPs map to the same contig and position", then yes, it is possible. If the two SNPs map to same contig location, but have different variation classes (e.g. a true SNP like "A/G", and an in/del SNP like "-/A"), we will not cluster them in the future. If the two SNPs have the same variation class (e.g. both are true single base substitutions), then we will merge them in a subsequent build. **(3/3/05)**

We observed two different SNPs at the same position, T/T and C/T, where the reference sequence is C/C. Would these two alleles be assigned individual refSNP numbers, or would they be assigned one refSNP number together?

We cluster submitted SNPs into a refSNP based on the submitted SNPs' genome mapping position, or based on their flanking sequence similarity. In your case, the two SNPs you found would be assigned one refSNP ID number. To report alleles in a submitted SNP, you would list all the alleles that you have observed in a mapping position. In your case, it would be "C" and "T", so you would report the SNP as "C/T". If you have individual sample genotypes, then you would report the genotype for each individual.

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**)

RefSNPs not Merging as Expected

Some SNPs for A2M seem redundant: rs3832852 and rs1799759 look like the same SNP, as do rs3832850 and rs35904656. rs5796338 and rs3080599 also look identical.

We believe that rs3832852 and rs1799759 in A2M are two separate refSNPs for the following reasons:

1. rs1799759 has variation as -/ACCAT, and rs3832852 has the variation as -/CCATA. If you put the variation in flanking context, the two different chromosome sequences are:

   rs1799759 (-/ACCAT)
   C-----AG
   CACCATAG

   rs3832852 (-/CCATA)
   CA-----G
   CACCATAG

   The deleted sequences above for the two refSNPs are shifted one base, so they remain separate SNPs. Currently in dbSNP, we do not have validation (freq or genotype) information for either of these two SNPs. If you have any validation information for either of these SNPs, please contact snp-sub@ncbi.nlm.nih.gov and submit your data to dbSNP.
2. rs3832850 and rs35904656 are 5 bases apart in mapping, and are therefore distinct SNPs.
3. Similarly, rs5796338 and rs3080599 are 16 bases apart in mapping, and are therefore distinct SNPs. (**10/5/07**)

Both rs6419492 and rs4601571 seem to describe the same SNP, yet both rs numbers exist separately in dbSNP. Am I missing something?

Yes, these two refSNPs should probably be merged. In general, we cluster submissions that co-locate on the genome, but our heuristics sometimes exclude particular cases. We prefer to err on the side of caution; better to leave a few co-located SNPs unclustered than cluster submissions which do not belong together. In this case, I can see from the submitter comments that the two refSNPs are from the same genomic location. Thanks for bringing this to our attention. **(3/13/05)**

Why is it that sometimes rs clusters do not merge as expected (e.g., rs1136410 and rs17853760)?

Due to processing timing constraints in b125, some rs numbers that map to the same positions were not merged, but will be merged in the next build. In future builds, SNPs that have different variation classes (and possibly different variation lengths) will not be merged even if they map to the same contig positions. (**2/9/06**)