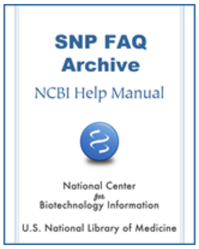
NLM Citation: SNP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. Merging RefSNP Numbers and RefSNP Clusters. 2005 Jul 7 [Updated 2014 Feb 18].
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Merging RefSNP Numbers and RefSNP Clusters

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

rs4823903, which has merged into rs4253690 (see RsMergeArch) still appears in SNPChrPosOnRef

Cases like rs4823903 have a bit of a twist in their history that you can see in SNPHistory:

```
 [502] MSSNP_LOAD.human_9606.2> select * from SNPHistory where snp_id =4823903
[502] MSSNP_LOAD.human_9606.3> go -mvertsnp_id:              4823903

create_time:         Feb 20 2003 01:26PM
last_updated_time:   Sep 07 2006 01:03PM
history_create_time: Oct 25 2006 12:19PM
comment: Re-activated:PHARMGKB:rs4823903|b126->rs60186231|b129:NT_011523.11_1851628
```

Submitter PHARMGKB submitted many high visibility SNPs in 2003, but also withdrew a large batch in 2006. Many of their SNPs are still valid and have been cited and used successfully in experiments by dbSNP users. PHARMGKB has since re-submitted many of their withdrawn SNPs to dbSNP, and we decided to "re-activate" the same rs numbers for those SNPs that have the same flanking sequences.

The "re-activation" was noted in the SNPHistory.comment field. For those re-activated SNPs that merged, I did not go back to change the merge history since it was an event that had indeed occurred. I can see that more work need to be done to clarify the multiple sequence of events for these SNPs. For users who actually use data from RsMergeArch and SNPHistory such clarification is necessary. (**09/09/08**)

The “rsCurrent" field in RsMergeArch contains 129,000 rs numbers not in the ASN.1 flat files or in a dbSNP web search. Is there a list of expired rs numbers?

Let me explain how we track both merged and deleted (an entirely different process from merging) refSNP (rs) numbers, by using a hypothetical example where a “chain merge” (multiple rs numbers merge into each other) occurs:

For example, let us say rs ”A” merged into rs ”B”, and later, rs ”B” merged into rs “C”. As a result of the first merge, the entry for rs ”A” in rsCurrent

is updated to rs “B”; after the second merge, rsCurrent is then updated to rs “C”. Now, if rs “C”’s submitters withdraw all the member ss numbers within the refSNP cluster rs “C”, then rs “C” will get an entry in the SNPHistory table (the SNPHistory table ONLY contains SNPs that have "become history" — that is, SNPs that have been completely deleted). Please see ftp file for SNPHistory.bcp (located in the snp/database/organism_data/species of interest directory). To find the column names for the SNPHistory table, download the human_9606_table.sql, which is located in the human organism_schema directory.

Getting back to RsMergeArch: since "withdrawing rs “C" is not a merge action, the table RsMergeArch is not updated. RsMergeArch is used to track "rs merge" actions only. I can see that this might be confusing, so when time allows, we will add the following explanation to the RsMergeAch table definition, to make the RsMergeArch.rsCurrent meaning clearer:

RsMergeArch is used to track each rs merge event.

If an rs number in RsMergeArch.rsCurrent is withdrawn from dbSNP by submitter request, then an the rs number of the same value as that in rsCurrent will be entered into the SNPHistory table (which contains deleted rs numbers only).

Please note: "rsCurrent" in RsMergeArch does not mean the "current rs number" in the current dbSNP build". (**08/12/08**)

Do you have tables that show those rs numbers that have changed between builds and those that have not?

RefSNP (rs) numbers will change only when a refSNP cluster merges with another cluster. When two clusters merge, the higher rs number is retired, and merged cluster takes the lower rs number.

You can find the dbSNP rs merge history in the RsMergeArch table, which is located in the organism_data directory for a particular organism in the dbSNP FTP site, or 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 refSNP page 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. (**12/03/07:11/03/08**)

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**)

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.

(**11/06/07:11/03/08**)

When some refSNP (rs) clusters merge, the higher number rs cluster is retired. Are retired rs numbers reused when new rs clusters are added?

When merging causes a rs number to retire, that rs number will never be reused. (**12/03/07**)

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.