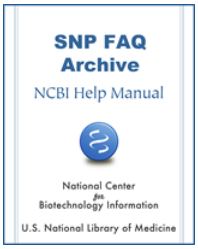
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: SNP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. Data Changes that Occur Between Builds. 2005 Jul 7 [Updated 2014 Feb 18].
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Data Changes that Occur Between Builds

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

New dbSNP Builds and Subsequent Changes to SNP Data

Changes to the dbSNP Schema Tables

I noticed that the coding non-synonymous function code has been subdivided. Can you list and define the subdivisions?

Function Code 41: "Nonsense" (coding nonsynonymous)

changes to the Stop codon

Function Code 42: "Missense" (coding nonsynonymous)

alters codon to make an altered amino acid in protein product

Function Code 44: "Frameshift" (coding nonsynonymous)

indel SNP causing frameshift

You can find up-to-date function codes and their definitions in the SnpFunctionCode.bcp.gz table located in the /shared_data directory of the dbSNP FTP site. (**10/27/08**)

I noticed that b127 SNPs are no longer associated with function class codes 1,2,5, and 7. Does dbSNP no longer use these function class codes?

As of build 127, function codes 1, 5 and 7 have been modified into two digit codes that will more precisely indicate the location of a SNP. The two digit codes have function codes 1, 5 or 7 as the first digit, each of these numbers keeping its original meaning, and 3 or 5 as the second digit, indicating whether the SNP is 3' or 5' to the region of interest. So the new function codes are as follows:

Function code13: "nearGene-3"

Where:
1=locus region
3= SNP is 3' to and 0.5kb away from gene

Function code 15: "nearGene-5"

Where:
1=locus region
5= SNP is 5' to and 2kb away from gene

Function code 53: "UTR-3"

Where:
5= UTR (untranslated region)
3= SNP located in the 3' untranslated region

Function code 55: "UTR-5"

Where:
5= UTR (untranslated region)
5(as the second digit)= SNP located in the 5' untranslated region

Function code 73: "splice-3"

Where:
7=splice site
3=3' acceptor dinucleotide

Function code 75: "splice-5"

Where:
7=splice site
5=5' donor dinucleotide

Function code 2, however, was retired permanently as of b127, as it identified a SNP as being in the coding region of a gene, but that other details about its location were unknown. Since mapping and annotation have improved dramatically since function code 2 was defined, it is no longer used. **(4/30/07)**

Changes in SNP Annotation

Why do the functional classifications for some variations change when a genome is re-assembled?

Functional annotation varies from build to build because the reference genome sequence, upon which SNPs are mapped, is itself changing from assembly to assembly. During each assembly, the algorithms used to define "genes" are refined to improve accuracy. Since gene features can be defined by various classes of evidence that vary in their certainty, fine-tuning is required in estimates of gene numbers and their precise exon structure on the genome. As an organism's annotation pipeline is developed, duplicate SNPs are identified and merged, spurious annotations are removed, and new evidence is included.

Depending on the organism, the net result of the alterations in assemblies and annotations mentioned above is that a SNP may be in an exon (or more generally, located in a gene) in one build, but then locate to an intron or UTR (or intergenic DNA) in the next build if the exon (gene) is removed between build releases. Once an organism's genome assembly comes closer to its finishing point, like that of the human genome, its annotation will become more stable, as will its SNP functional classification. (**04/05/06**)

Changes in Mapping

Where can I find the mapping table changes for SNP positional data on genomes and GenBank sequence records?

Please see the build 125 mapping doc, located in the "specs" subdirectory of the SNP FTP directory. **(10/03/05)**

Why is the physical position of all of the SNPs in B125 less than in previous builds?

Build 125 mapping tables are now 0 based — that is, the position of the first base is 0 instead of 1 as in previous builds. For other mapping table changes, please see the build 125 mapping doc, located in the "specs" subdirectory of the SNP FTP directory. **(10/04/05)**

Since b125 SNPs are mapped to both the NCBI reference assembly and to the Celera assembly, if an rs number is located on both assemblies, will they in the same position? Which assembly should we use?

The NCBI reference and the Celera assembles will differ depending on the region, but for any one region, one assembly may be more accurate than the other. Please note that the Celera assembly is not updated with new

build cycles, and is therefore essentially static. The NCBI reference assembly is updated periodically. (**1/12/06**)

Is it possible for the alleles and the orientation of a refSNP flanking sequence, change either between builds, or upon a new build of dbSNP (e.g. rs17871378: could aacaaaggct [A/C] acggaaggag change to ctccttccgt[G/T]agcctttgtt)?

The case you describe should never happen. (**9/22/06**)

Individual refSNP Mapping Changes

I found rs123456 in the current build and was wondering if it will be valid and point to the same SNP in future builds.

Approximately 8% percent of the rs numbers in dbSNP have been retired since the inception of dbSNP, so you could say that rs numbers are not entirely stable. Any rs number in dbSNP could retire for one of the following reasons:

1. rs numbers occasionally merge with other rs numbers because they are found to map to the same location on the genome. When such a merge happens, ss numbers of the higher rs number are reassigned to the lower rs number, and the higher rs number is retired.
2. If all ss numbers in an rs cluster are withdrawn, then the corresponding rs number is retired.
3. We break an existing cluster and re-instantiate a retired rs number based on a reported conflict from a dbSNP user (a rare occurrence).

rs2994917 had two positions in B124 , but shows only one position in B125. What changed?

A reference SNP, or rs, is a cluster of submitted SNPs (ss). Initially, each ss is mapped to the current genome assembly and clustered to an existing rs. The longest ss in a rs cluster is used for mapping to subsequent genome assemblies.

There are a few of reasons why an rs may change between builds:

1. A constituent ss of the rs cluster is withdrawn by the submitter.
2. A new ss is added to the rs cluster. This category includes rs merges, where two previously independent refSNPs become one (the lowest rs number is assigned to the merged clusters in a rs merge)
3. The Mapping algorithm is continuously being improved.

In the case of rs2994917, reasons 1 & 2 can be ruled out, since an ss was neither withdrawn, nor a new ss added between the builds. The reason for the change in rs2994917 was alterations in the algorithm between B124 and B125. These alterations allowed hits that just passed the threshold of a "good hit" in build 124 to appear below this threshold in B125. **(11/23/05)**

Why are SNPs submitted by HGBase that mapped to the genome assembly in previous builds of dbSNP no longer mapped? For example, rs1800324 previously mapped to chromosome X, but is no longer.

It is true that rs1800324, which mapped to the genome assembly in previous releases of dbSNP, no longer maps. The reason is that the internal matching heuristics for NCBI's MegaBLAST program, which is used for mapping SNPs, has been changing, and this has resulted in mapping changes.

Because of the new matching heuristics, refSNPs with very short flanking sequences can be dropped from the map. rs1800324, submitted by HGBASE, has very short flanking sequences (25 bp on each side), and this may have contributed to it being dropped from the map in b125. If other HGBase submissions also have very short flanking sequences, they may also have been dropped for the same reason. (**2/3/06**)

Why does rs3823342 have two different chromosome positions?

Starting with b125, the SNPs in dbSNP will be mapped to both the NCBI reference assembly and the Celera assembly. Because rs3823342 is mapped to both the NCBI reference assembly and the Celera assembly, the two different chromosome positions you found are for the two different assemblies. You can see this clearly in the Integrated Maps section of the rs3823342 report. **(1/3/06)**

Why does rs3128991 have two different local loci?

Starting with b125, the SNPs in dbSNP will be mapped to both the NCBI reference assembly and the Celera assembly. Because rs3128991is mapped to both the NCBI reference assembly and the Celera assembly, the two different local loci you found are for the two different assemblies. You can see this clearly in the Gene view section of the rs3128991 report. **(1/3/06)**

Changes to RS numbers?

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 now 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.

You can also 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, . (**12/03/07:11/03/08**)

Changes in Numbers of SNPs Reported

I find that 20% of the refSNP (rs) numbers in Build 129 cannot be found in Build 130. 20% seems rather high. Is this correct?

20% of b129 rs numbers have been merged in b130. And you're right, this percentage is rather high, and is the result of inappropriate rs number assignment to non-novel SNPs in several large submissions that contained only a small percentage (10-20%) of novel SNPs. We later corrected this mistake by merging the non-novel rs numbers from these submissions with existing rs numbers. (**06/15/09**)

Are new SNPs added to dbSNP between dbSNP builds?

In general, new SNPs are added to dbSNP only during a build. If you want your local copy of dbSNP to stay current, it is necessary to update your local database every time dbSNP releases a new build.

There might be, however, some rare instances when a new dbSNP build has no new SNPs—just new frequency data, genotype data, or new mapping information. When dbSNP announces a new build, check the snp_summary page to see details on the content of that build.

Why is it that when I searched with human genome on dbSNP three weeks ago, the search retrieved 2456 records, but when I performed the same search today, it retrieved only 2127 records?

The human genome is still undergoing occasional changes between builds. This may have caused some SNPs to recluster to different locations on the genome during dbSNP build 121, and/or they have been annotated with a different functional class.

Why is there a large discrepancy between the numbers of submitted mouse SNPs(ss) and refSNPs(rs) reported for build 125 (b125) on the dbSNP summary page, and the numbers of b125 mouse ss and rs that can be loaded from the dbSNP FTP files?

The numbers of submitted SNPs and refSNPs reported on the summary page for b125 are correct. Around the time build 125 (b125) was released to the public, PERLEGEN submitted millions of mouse SNPs to dbSNP. These submissions did not get clustered, however, until build 126, so the number of submitted SNPs was over 6 million in the b125 summary, while the total number of refSNPs for the same build was only about 1.8 million. I suggest that you download the b126 FTP files since they contain three to four times more mouse SNPs than do the b125 files.

For future reference, you can see the number of SNPs submitted for a particular organism and the date the submissions were made by going to the dbSNP Summary page, and clicking on the number of new submission (ss#s) for a particular organism. (**4/6/06**)

New Genome Builds and Subsequent Changes to Linked NCBI Resources

Between genome builds, are the data in Entrez Gene and dbSNP static or frequently updated?

Public dbSNP data are updated only at build release time, which is roughly every 1 to 2 months. For more details about Entrez Gene, email the NCBI helpdesk: info@ncbi.nlm.nih.gov

When a new genome build is released and new contig gi numbers are created, is it coordinated across Entrez Gene, dbSNP, and any other databases at NCBI that might be affected by the changes?

Yes, we update our internal databases.

Are some of the contigs, or chromosome segments, consolidated or revised and given new gi numbers when NCBI releases a new genome build?

Yes.