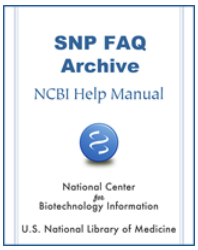
U.S. National Library of Medicine
National Center for Biotechnology Information

**NLM Citation:** SNP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. Discrepancies between Cluster Report Data and FTP Data Contents. 2005 Jul 7 [Updated 2014 Feb 18].
**Bookshelf URL:** https://www.ncbi.nlm.nih.gov/books/




# Discrepancies between Cluster Report Data and FTP Data Contents

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

**rs4646149 shows the submission of ss69370884 in build 127 on the website, but there's no record of that submission in the dbSNP FTP SubSNP.bcp file.**

ss69370884 was submitted on March 22, 2007, but the SubSNP.bcp file on FTP site was created on March 6, 2007. Build 127 was released on March 8, 2007.

In general, new submitted SNPs (ss) are made public more frequently than there are SNP build releases so submitters can see their submissions more quickly(without having to wait months for a new build release). In the past, new submitted SNPs were not clustered until the new build.

Beginning in May 2007, however, we cluster the new submitted SNPs as they arrive, so when these submitted SNPs (including ss69370884) are made public they also show up in their clusters, as you have observed. All the table .bpc files, however, are created only once per build, so that is why you don't see this submitted SNP in the .bcp files. (**7/2/07**)

**Why do web queries of rs939820, rs10205833, and rs7597158 return sequences that do not match the exemplar sequences for these SNPs found using a database query?**

When you refer to the sequences as ”not matching”, I assume that you are referring the fact that they don’t match at the point of variation, since in rs939820, for example, the two flanking sequences are the same with the exception of the point of variation.

On the RefSNP (rs) page, we show the rs FASTA using the IUPAC code for variations, while on the submitted SNP (ss) page, we show the ss fasta using the submitted observed sequence. In most cases, all member ss of an rs cluster have the same allelic states in the same orientation, so the rs variation matches the ss exemplar variation. There are cases, however, where the rs variation does not match the ss exemplar variation. For example, if an ss exemplar has an A/G variation, and another ss from the cluster in the same orientation has an A/T variation, then the rs allele list will read A/G/T since it includes all member ss alleles. If you viewed the rs allele list converted into IUPAC code (remember the refSNP page shows the flanking sequence in IUPAC), it would show a D representing A or G or T.

Most of the submitted SNPs have the same variations in the rs clusters you mentioned, but one or two of the ss in each cluster have an extra allele. All of the ss in the rs939820 cluster have an A/G variation, with the exception of one ss that has an -/A/G variation. All the ss in the rs10205833 cluster have a C/G variation, with the exception of one ss that has a C/G/T variation. Most of the ss in the rs7597158 cluster have an A/G variation, while the ss exemplar has a -/G variation. In these cases, the refSNP page variation list includes all allelic states: for rs939820, instead of an R, it is N; For rs10205833, instead of an S, it is B that represents for C or G or T; for rs7597158, since the ss exemplar has a -/G variation, while most of the other ss in this cluster have an A/G variation, the refSNP FASTA shows an N at the variation point.

I will update dbSNP's variation representation for "mixed variation" clusters to include all the allele lists from all the submitted SNPs.(**8/18/06**)