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

SNP FAQ
Archive

NCBI Help Manual

National Center
for
Biotechnology Information

U.S. National Library of Medicine

Compact SNP data: The Bitfield

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

I work with many thousands of SNPs, and would like to visualize information on a large scale. Do you have a format that will hold a lot of data for each SNP in a compact fashion?

SNP has its data available for many organisms as a bitfield, which can compactly hold a great deal of information. For humans, you can access SNP_bitfield.bcp.gz in the human subdirectory of the organism_data directory. The specs for SNP_bitfield located in the specs directory of the dbSNP FTP site.

The bit field can be retrieved using the NCBI C++ Toolkit with its feature iterator and annotation selection classes. This would imply that you have developed a C++ program that will read the bit field too. The bit field can also be retrieved from raw XML or ASN.1 dumps taken from Entrez. You can use the PERL programming language to extract the bit field from there without using C++. (**11/08/07**)

I downloaded SNP_bitfield.bcp.gz and SubInd_ch18.bcp.gz to get rs numbers, locations, frequencies, and flank for SNPs on chromosome 18, but found only numbers.

The only way to read the bitfield is to have retrieved it using the NCBI C++ Toolkit with its feature iterator and annotation selection classes as mentioned in the second paragraph of the previous FAQ in this section. This would imply that you have developed a C++ program that will read the bit field too. The bit field can also be retrieved from raw XML or ASN.1 dumps taken from Entrez. You can use the PERL programming language to extract the bit field from there without using C++.

To get "RS numbers, locations, frequencies, and flanking sequence for all SNPs on human chromosome 18", you will need to download two files:

- You will need the genotype report file by chromosome (gt_chr18.xml.gz) located in the /human_9606/genotype/ directory of the dbSNP FTP site. In this file you will find the rs number, chromosome positions and allele frequencies (organized by each member ss and population).
- The best file to download for refSNP flanking sequences is the rs_ch18.fas.gz file located in the /human_9606/rs_fasta/directory of the dbSNP FTP site.

(**09/10/08**)