



Ancestral Allele Data

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

Is dbSNP's "ancestral allele" the same as HapMap's "reference allele"?

The NCBI "Reference allele" for a given SNP refers to the nucleotide base on the NCBI reference assembly at the SNP's position. We think that this is what HapMap's "reference allele" refers to as well. We can confirm this if you will send us a few examples of where HapMap uses the term "reference allele".

The dbSNP refSNP report page does not directly use the term "reference allele". In the "Integrated Maps" section of the refSNP cluster report, there is a column called "Contig Allele" since "Contig Allele" is the allele on the contig at the SNP's position. The "reference allele" is the "contig allele" value when the assembly type is "reference" in the map table. (**Please Note:** a reference contig's name starts with "NT_" and a Celera contig's name starts with "NW_".) The contig alleles on different assemblies are often the same, but not always.

[Rs3000](#) is an example of a SNP that has different contig alleles.

The ancestral alleles provided in a human refSNP cluster report is the SNP allele as found in the chimpanzee; I should mention here that not all SNPs in dbSNP have ancestral allele data. This data was provided by Dr. Jim Mullikin at NHGRI. There is more information about ancestral alleles in the ancestral allele subsection of the Data Content section of this archive. (01/04/08)

How do I find ancestral allele information in dbSNP?

You can get these data from two the tab-delimited table dump files on the [dbSNP FTP site](#): SNPAllele.bcp and Allele.bcp. When the SNPAllele.ancestral_flag is 1, the allele_id is an ancestral allele. To get the actual allele, please use the allele table (Allele.allele_id and Allele.allele). The DDL for shared tables is located in the [shared_schema](#) subdirectory, and the organism table DDL is located in the [organism sub-directory](#) (for this example I'm using human as the example organism).

You'll find that the rsID in "SNPAncstralAllele", is stored as an integer in the first column of the table. Here's an example to help you — say you want to find the ancestral allele for rs3:

1. Look for the number "3" in first column of the SNPAncstralAllele table and you'll find that the corresponding ancestral allele ID (located in the second column) is 7.
2. Now, look in the Allele table for the allele letter that corresponds to Ancestral allele ID = 7. In the allele table, the allele ID number is again located in the first column:

```
7 C 2003-02-22 01:11:00.0 2 2003-10-06 17:42:00.0
```

So you can see from this example, that the ancestral allele for rs3 is "C". (03/08:08/08)

How do I download a flat file that contains the ancestral state of each SNP (when this ancestral state is known)?

You can get the ancestral state of each SNP (when this ancestral state is known) by using the following two tables from the FTP dump files:

1. Go to the [organism_shared_data](#) file located in the dbSNP FTP site, and download the Allele.bcp.gz file.
2. Go to your organism's (human, in this case) "[organism data](#)" file located in the dbSNP ftp site and download the "SNPAncstralAllele.bcp.gz" file.
3. You can get the column definitions for the SNPAncstralAllele and Allele.bcp.gz tables from the [dbSNP main table.sql.gz](#), which is located in the shared_schema file of the dbSNP ftp site. (11/16/07)

Determining Ancestral Alleles

How was the SNPAncstral Allele table derived?

SubSNP.ancestral_allele is a table for storing ancestral alleles submitted to dbSNP. The "Ancestral allele" field in the dbSNP submission form is an optional field, so very few submissions have this information. That is why the Ancestral Allele field is mostly empty in SubSNP table. The SNPAncstralAllele table has data from only a single source — Dr. Jim Mullikin of the National Human Genome Research Institute (NHGRI). Dr. Mullikin derived his ancestral allele data by the comparison of human DNA to chimpanzee DNA, and his methodology is available online at [PLoS](#). The last time Dr. Mullikin provided us with ancestral data was in May 2004. (6/13/07)

Could you please define the ancestor upon which SNP ancestral allele information is based?

The National Human Genome Research Institute (NHGRI) determines the ancestral alleles for human SNPs by the comparison of human DNA to chimpanzee DNA. The methodology used for finding these ancestral alleles is available online at [PLoS](#). (6/13/07)

How do I determine if a SNP is an ancestral allele?

Human SNP ancestral alleles are determined by comparison with primate DNA, so in general, they're based on chimpanzee sequence.

You can get ancestral allele data from dbSNP in a couple of ways:

1. If the ancestral state of the SNP in which you are interested is known, and is contained in dbSNP, look in the "Allele" section of the the RefSNP page, which is located at the top right hand side of the report.
2. You can also get the ancestral state of a SNP (when this ancestral state is known) by using the following 2 tables from the FTP dump files:
 - a. Go to the [organism_shared_data](#) file located in the dbSNP FTP site, and download the Allele.bcp.gz file.
 - b. Go to your organism's (human, in this case) "[organism data](#)" file located in the dbSNP ftp site and download the "SNPAncstralAllele.bcp.gz" file.
 - c. You can get the column definitions for the SNPAncstralAllele and Allele.bcp.gz tables from the [dbSNP main table.sql.gz](#), which is located in the shared_schema file of the dbSNP ftp site. (2/7/05)

I want the ancestral allele for a number of rsIDs, but can't find rsID numbers in Allele.bcp.gz and SNPAncstralAllele.bcp.gz.

If you go to the dbSNP Data Dictionary and search for "[SNPAncstralAllele](#)", you'll find that the rsID is stored as an integer in the first column of the table.

Here's an example to help you — say you want to find the ancestral allele for rs3:

Look for the number "3" in first column of the SNPAncstralAllele table and you'll find that the corresponding ancestral allele ID (located in the second column) is 7.

Now, look in the Allele table for the allele letter that corresponds to Ancestral allele ID = 7. In the allele table, the allele ID number is again located in the first column:

```
7 C 2003-02-22 01:11:00.0 2 2003-10-06 17:42:00.0
```

So you can see from this example, that the ancestral allele for rs3 is "C". (08/29/08)

Allele.bcp.gz and SNPAncestralAllele.bcp.gz show the ancestral allele for rs10465407 is "A", but my multi-genome alignment shows "G" is ancestral.

The method used to derive the ancestral alleles that are stored in SNPAncestralAllele may be different from your method or possibly outdated as more sequence evidence becomes available suggesting an alternate allele.

Please see the FAQ "How was the SNPAncestralAllele table derived?" (08/29/08)

Annotating Ancestral Alleles

Is there a program that I can use which will allow me to annotate ancestral alleles to existing SNPs? Would dbSNP accept submissions of such data?

Ancestral allele data is provided to dbSNP as part of the SNP assay data that is provided in dbSNP submissions, so we don't have a program that allows us to annotate ancestral alleles, and we don't currently have a mechanism for third party annotation of SNPs. (3/29/05)

Interpreting Ancestral Allele Field in refSNP Report

What does the "?" in the ancestral allele field in a refSNP report mean?

"?" in the ancestral allele field indicates that we do not know the ancestral allele for the SNP. (3/3/05)