dbSNP Reference SNP (rs) Stand Orientation Reporting Updates

Date: Jan 31, 2019

From the original public release in 1998 until 2018 (dbSNP Build 151), the alleles associated with a dbSNP RefSNP were maintained in the flanking sequence orientation of the first submitted SNP of that cluster. As the human genome assembly became available in 2001, each RefSNP was mapped to the assembly, but the allele orientation was preserved to be the same as the flanking sequence of the first submitted SNP of that cluster. Therefore, some RefSNPs were reported on the forward (FWD) strand of the assembly, while others were reported on the reverse (REV) strand. As the assembly model evolved (and now at GRCh38/hg19), some portions of the assembly sequence swapped orientation, and the FWD/REV attribute of the RefSNP changed accordingly. As a result, a RefSNP could be FWD to one genome assembly version and REV to another version. However, since a RefSNP was defined by the flanking sequence of the first submitted sequence, the strand for the RefSNP remained the same. This strand or orientation stability was critical to interpret the meaning of the alleles, independent of the assembly version.

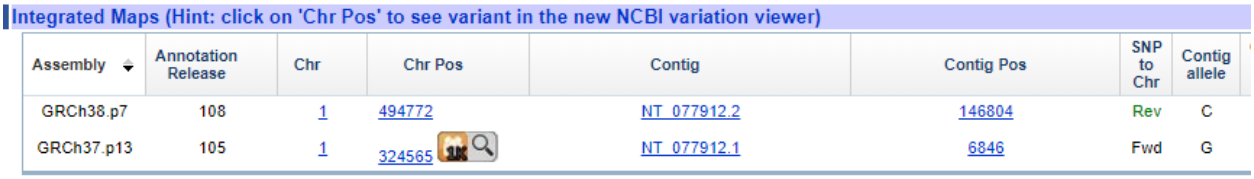
To better manage the exponential growth in submitted and aggregated variants, dbSNP has adopted a new architecture and data products. dbSNP 2.0 Build 152 was officially released to public at the end of 2018. In the new RefSNP report, the allele orientation is _always_ forward to whichever sequence it is reporting on, and the alleles used for identifying the RefSNP type are on the preferred top-level sequence (typically GRCh38/hg19). This is aligned to how variants are reported across the community in VCF and HGVS formats. It is also aligned with NCBI's new internal format, SPDI (https://doi.org/10.1101/537449) and eliminates ambiguity in reporting across resources. As a result, the orientation of the alleles relative to the sequence is unnecessary, since all variants are FWD to the particular sequence they are reported upon.

Even though reporting RefSNP alleles on genome assembly orientation has its advantages, we are also aware that it can be challenging for users to map the older RefSNP and allele orientation model to the new reporting. Therefore, we have prepared the VCF files for rs# which were formerly reported as REV for either GRCh37 or GRCh38 and would be inconsistent with the current reporting standards.

For example, rs750036001 is a C/G variant created in 2015 from ExAC submission: EVA_EXAC|EXAC_0.3.1:g324565g>c. with the following flanking sequence:

5' -> 3': cgtcacggcagcctccgcagatgag [G>C] ctactgcctcacaacagcctccaca

This is in the FWD strand of GRh37, but REV to GRCh38, as illustrated in the Classic SNP RefSNP report, below:



Integrated Maps (Hint: click on 'Chr Pos' to see variant in the new NCBI variation viewer)

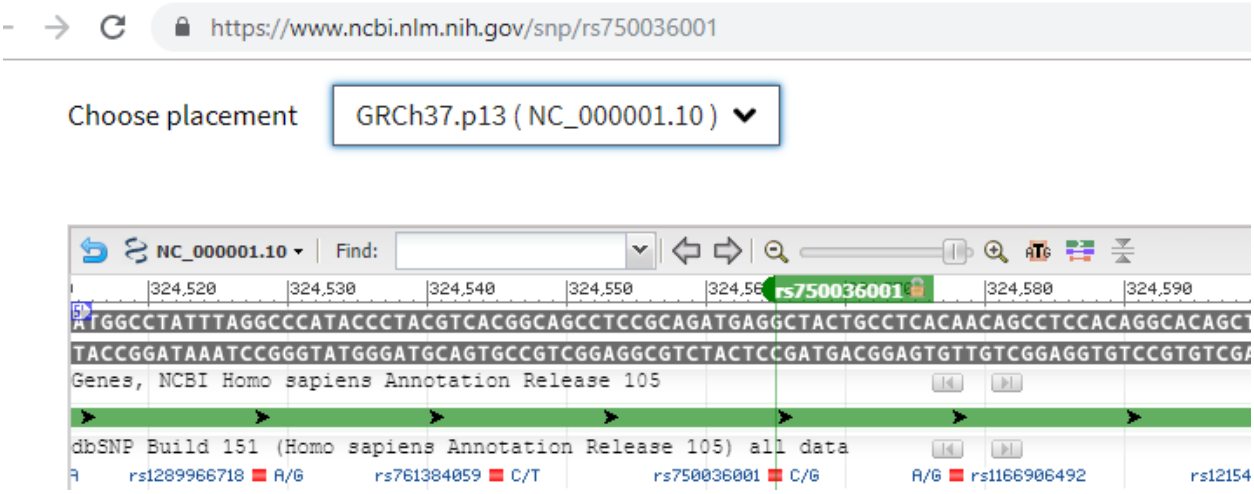
Assembly	Annotation Release	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele
GRCh38.p7	108	1	494772	NT_077912.2	146804	Rev	C
GRCh37.p13	105	1	324565	NT_077912.1	6846	Fwd	G

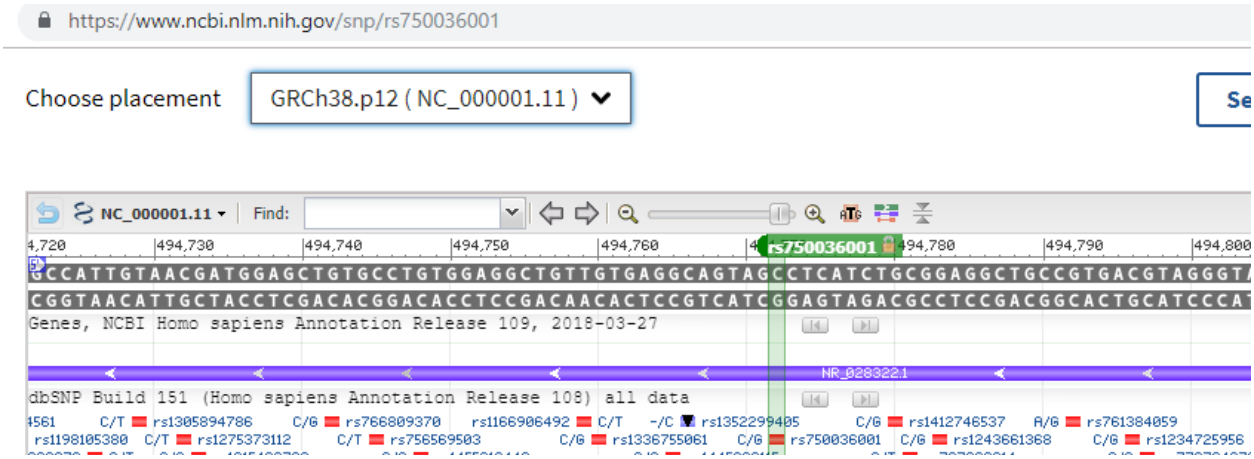
In the new RefSNP report, the fact that GRCh37 and GRCh38 has flipped orientation in the region surrounding this rs750036001 can been seen in the “Genomic view” section at the bottom of the RefSNP page using “Choose placement” to choose GRCh37 and GRCh38 and the result is illustrated below.

on GRCh37 5’ to 3’: ATGAG[G>C]CTACT ,

on GRCh38 5’ to 3’: AGTAG[C>G]CTCAT

Notice the above are reverse complemented.





The fact that the assembly sequence has flipped orientation in the region flaking RefSNP rs750036001 can also be seen in the new RefSNP Web report “Genomic Placements” section as illustrated below. Notice the ref and alt alleles are “reversed complemented”:

Genomic Placements

Sequence name	Change
GRCh37.p13 chr 1	NC_000001.10:g.324565G>C
GRCh38.p12 chr 1	NC_000001.11:g.494772C>G

In the two vcf files are here https://ftp.ncbi.nih.gov/snp/redesign/archive/b152/VCF/supplement/:

- rs.with.flipped.orient.grch37.vcf
- rs.with.flipped.orient.grch38.vcf

You will see rs750036001 reported below clarifying the RefSNP orientation to different assemblies (white space added here for clarity):

In rs.with.flipped.orient.grch37.vcf :

NC_000001.10 324565 rs750036001 G C . .

FREQ=ExAC:0.9707,0.02929|GnomAD:0.9992,0.0008347;

RS2GRCh37Orien=0;

RS2GRCh38Orien=1;

HGVS_GRCh37=NC_000001.10:g.324565G>C;

HGVS_GRCh38=NC_000001.11:g.494772C>G

The same rs in rs.with.flipped.orient.grch38.vcf:

NC_000001.11 494772 rs750036001 C G . .

FREQ=ExAC:0.9707,0.02929|GnomAD:0.9992,0.0008347;

RS2GRCh37Orien=0;

RS2GRCh38Orien=1;

HGVS_GRCh37=NC_000001.10:g.324565G>C;

HGVS_GRCh38=NC_000001.11:g.494772C>G

We hope that the VCF files will help you transition your RefSNP reporting system to consume data from the new RefSNP reports. We highly appreciate your feedbacks. Please send us your feedback, comments, and suggestions to dbSNPredesign@ncbi.nlm.nih.gov.