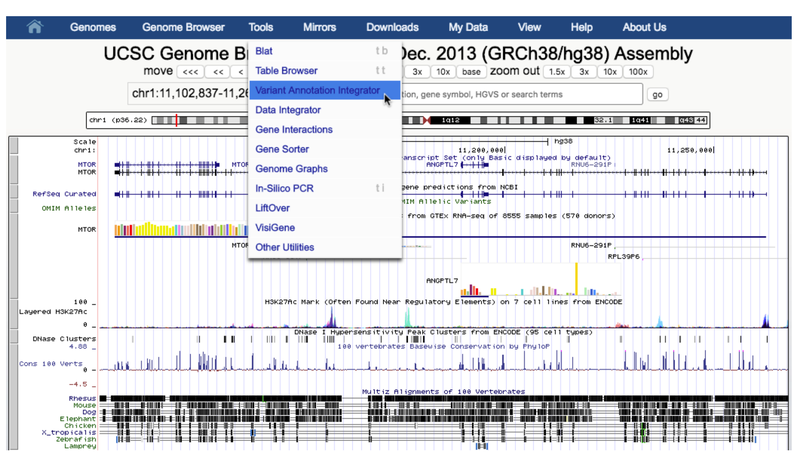
# HGVS variant nomenclature in the UCSC Genome Browser

Angie S. Hinrichs, Christopher M. Lee, Christopher Villarreal, Robert M. Kuhn, Maximilian Haeussler, Brian T. Lee, Luvina Guruvadoo, Cath Tyner, Ann S. Zweig, W. James Kent
University of California Santa Cruz Genomics Institute

## UCSC Genome Browser and Tools

The UCSC Genome Browser is a powerful web-based visualization tool for public genomic datasets as well as custom user-provided data and annotations. In addition to the graphical display of the Genome Browser, there are several tools for rapidly aligning sequences to the genome, adding one's own data, mining the data, and sharing customized views.




## HGVS variant nomenclature

The Human Genome Variation Society (HGVS) publishes recommendations for a consistent and precise representation of sequence variants [den Dunnen *et al.* 2016]. Many journals require that variant descriptions follow HGVS recommendations and HGVS terms are often used in clinical reports. The general form of an HGVS term is

*{accession}*:*{type}*.*{position}{change}*

For example, the following terms all describe ClinVar variant RCV000272467 (in gene PLEKHG5):

| | |
|---|---|
| NC_000001.10:g.6537598G>T | genomic variant (GRCh37/hg19) |
| NC_000001.11:g.6477538G>T | genomic variant (GRCh38/hg38) |
| LRG_262:g.47472C>A | genomic variant (Locus Reference Genomic) |
| NM_020631.4:c.34C>A | coding variant |
| LRG_262t1:c.34C>A | coding variant |
| NP_065682.2:p.Pro12Thr | protein variant |
| LRG_262p1:p.Pro12Thr | protein variant |

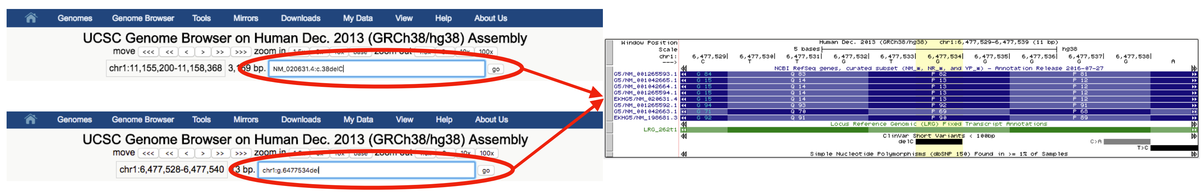
The recommendations cover insertions, deletions, and multi-base changes as well as intronic changes and UTR changes in coding genes, for example:

| | |
|---|---|
| NC_000016.10(NR_003501.1):n.2465-39_2465-37dup | duplication in noncoding intron (correct HGVS) |
| NR_003501.1:n.2465-39_2465-37dup | duplication in noncoding intron (common practice) |
| NM_153818.1:c.-52_-51del | deletion in 5′ UTR |
| NM_015074.3:c.*279_*280insGT | insertion in 3′UTR |
| NC_000001.11:g.6477538_6477539delinsTT | 2-base substitution in genome |

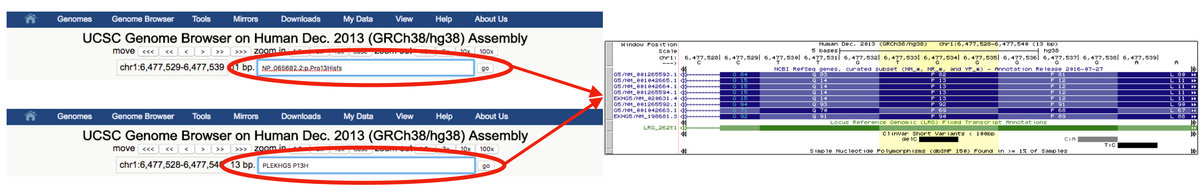
## Mapping HGVS (and HGVS-ish) terms to genomic locations

The UCSC Genome Browser accepts HGVS terms and similar notations as position/search inputs for navigating to a variant's genomic position with several extra bases on either side for context.

### Nucleotide-level




### Amino acid-level




## References:

den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux A, Smith T, Stylianos E. Antonarakis, Peter E.M. Taschner. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat. 2016 Jun;37(6):564-9.

Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, Fischer CM, Gibson D, Gonzalez JN, Guruvadoo L, *et al*. The UCSC Genome Browser database: 2017 update. Nucleic Acids Res. 2017 Jan 4;45(D1):D626-D634.

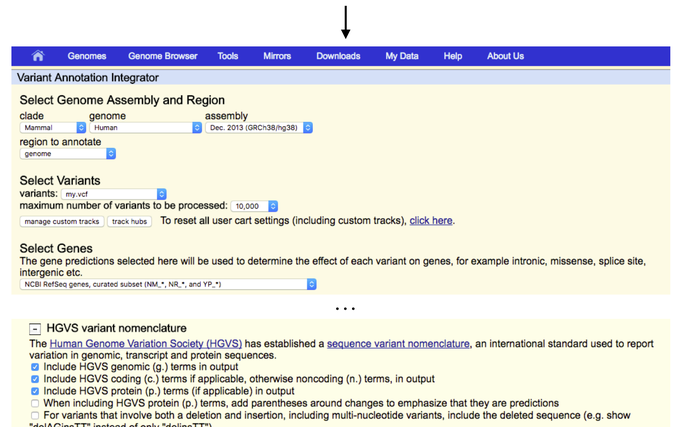
## Acknowledgements:

This work was funded by the National Human Genome Research Institute (5 U41 HG002371 to UCSC Center for Genomic Science). We would like to acknowledge the work of the UCSC Genome Bioinformatics technical staff (http://genome.ucsc.edu/staff.html), our many collaborators, and our users for their feedback and support.

## UCSC Variant Annotation Integrator and vai.pl

The Variant Annotation Integrator (VAI) provides an easy way to add annotations to user-provided variants (for example, from a VCF custom track) including functional effect on transcripts, various impact scores and more. When RefSeq transcripts are selected, HGVS variant nomenclature terms can be added to the output as well.

```
##fileformat=VCFv4.0
#CHROM  POS      ID         REF    ALT    QUAL    FILTER  INFO
1       6477531  rs982875622  T      A      .       .       .
…
```




Since VAI is a web tool, it can handle only a limited number of variants without timing out, and using VAI on the UCSC Genome Browser website requires transmitting variant data over the Internet, which is not safe for confidential human subject data.

However, we now have a command-line script for running VAI on your own computer with local variant files: vai.pl. The easiest way to get up and running with vai.pl is to install our Genome Browser in a Box (GBiB) or Genome Browser in the Cloud (GBiC).

```
# vai.pl hg38 my.vcf –geneTrack=ncbiRefSeqCurated
```

```
## ENSEMBL VARIANT EFFECT PREDICTOR format (UCSC Variant Annotation Integrator)
## Output produced at 2017-10-20 14:55:04
## Connected to UCSC database hg38
## Variants: my.vcf
## Transcripts: NCBI RefSeq genes, curated subset (NM_*, NR_*, and YP_*) (hg38.ncbiRefSeqCurated)
Uploaded Variation  Location  Allele  Gene Feature  Feature type  Consequence  Position in cDNA  Position in CDS  Position in protein Amino acid change
    Codon change  Co-located Variation  Extra
rs982875622  chr1:6477531  A  PLEKHG5  NM_001265593.1 Transcript  missense_variant  274  248  83  Q/L  cAa/cTa  -  HGVSG=NC_000001.11:g.
6477531T>A;HGVSCN=NM_001265593.1:c.248A>T;HGVSP=NM_001265593.1:p.Gln83Leu;EXON=2/21
…
```

## Mapping genomic variants to transcript HGVS terms

Converting a genomic variant call, typically from a VCF file, to HGVS genomic, transcript and protein terms is straightforward for the majority of cases, in which the reference genome and RefSeq or LRG transcript align perfectly with no mismatching or skipped bases. Things become increasingly complicated when the variant call has more than one valid placement within a repetitive region and when the reference transcript sequence differs from the reference genome sequence.

**Ambiguous indel mapping: shifting conventions**

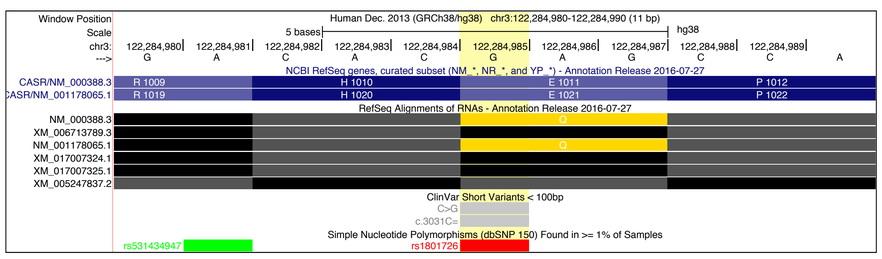
`AAAAA/AA`: which bases were deleted?

| left/5′-shifted: | right/3′-shifted: |
|---|---|
| `ref TAAAAAG` | `ref TAAAAAG` |
| `alt T---AAG` | `alt TAA---G` |

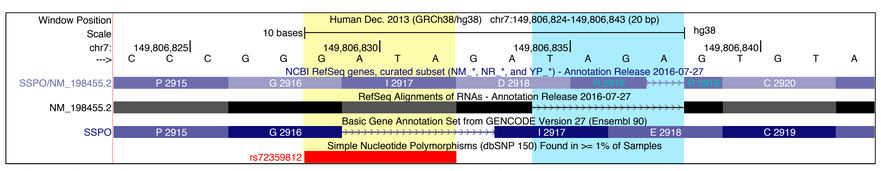
<<< VCF: left/5′-shifted on genome + strand
>>> HGVS: right/3′-shifted on strand of reference {genome, transcript, protein} – genome/transcript may differ!

**Disagreements between reference genome and RefSeq transcripts**

RefSeq transcript sequences usually align perfectly to the reference genome, but since the reference genome is a mosaic of individual genomes that include rare or singleton variants (and a few errors), thousands of transcript alignments contain mismatching bases. Hundreds even have insertions/deletions in which the reference genome is missing a few bases or has extra bases relative to the transcript. In these cases, a genomic variant call might mean no change to the transcript. Worse yet, the lack of a genomic variant might mean a damaging change to the transcript.




Above: Base mismatch between RefSeq transcript NM_000388.3 (C) and GRCh38 (G). The RefSeq Alignments track highlights the mismatching amino acid translation (RefSeq CAG/Q, GRCh38 GAG/E). dbSNP variant rs1801726 has a minor allele frequency of 5.6% for G. ClinVar annotates both alleles, the GRCh38/minor allele NM_000388.3:c.3031C>G and the RefSeq transcript/major allele NM_000388.3:c.3031C=, as Benign.




Above: This 20-base region in GRCh38 aligns to a 16-base region of RefSeq transcript NM_198455.2. rs72359812 has a minor allele frequency of 27.6% for the RefSeq transcript allele. Due to falling in a repetitive region, the 4 extra bases in GRCh38 are aligned to different locations by different algorithms and create an alignment gap that is easy for software to misinterpret as an intron. A variant call of a 4-base deletion from the reference actually means no change to the RefSeq transcript; the GRCh38/major allele contains a frameshift that induces an early stop codon. ClinVar does not annotate this variant.

## More information

- Search for answers in our mail list archives: http://genome.ucsc.edu/contacts.html
- Email us a new question: genome@soe.ucsc.edu
- UCSC training (workshops, videos, tutorials): http://genome.ucsc.edu/training/
- Blog: http://genome.ucsc.edu/blog/