NC\_000011.9:g.534286C>A

Workflow to review variant with HGVS representation:

1. Look up accession and version number on NCBI database if you have not memorized which reference genome is for which accession and version numbers.
   1. <https://www.ncbi.nlm.nih.gov/assembly/help/> > Search: NC\_000011.9 > Select “See NC\_000011.9”
   2. NC\_000011.9 is for GRCh37.p13 Primary Assembly
   3. Ideally it would be better to have access to @ Genome assembly database <https://www.ebi.ac.uk/ena/browse/genome-assembly-database> in which you could use my quick Github [script](https://github.com/BJWiley233/Practical-Computer-Concepts-Files/blob/master/linux_shell/variant/get_assemblies.sh) to draft a query to the database for instance if you had a list of accession.numbers for Human you could search for the assemblies.
2. Obtain Gene Name and Protein information. There are various ways to do this. Here are two options interface options and one program option.
   1. (Preferred) UCSC Genome Browser @ <https://genome.ucsc.edu/cgi-bin/hgGateway?hgsid=818174999_fMu2exunJ89SYjUA8e1bovaUiZED>
      1. You can search directly in the ‘Position/Search Term’ search field with the HGVS representation NC\_000011.9:g.534286C>A
      2. Alternatively, under ‘Human Assembly’ you can select GRCh37/hg19 > Go > Enter chromosome and position ‘chr11:534286-534286’> go
      3. Here we can use the interface to obtain the information needed.
         1. Under ‘UCSC Genes’ we will see the Gene on the left and whether there are alternate transcripts. By clicking each one of the alternate transcripts for example [transcript variant 2](https://genome.ucsc.edu/cgi-bin/hgGene?hgg_gene=uc010qvw.2&hgg_prot=uc010qvw.2&hgg_chrom=chr11&hgg_start=532241&hgg_end=535550&hgg_type=knownGene&db=hg19&hgsid=818179607_rLwHqaVTIRbJZDrtfHFB82mla5mo), we can see this region of the genome codes for the HRas proto-oncogene, GTPase gene. This gene belongs to the Ras oncogene family. The GTPase family binds GTP and hydrolyzes it to GDP+P to perform their function. We can also find more information with each transcript regarding the total exon counts, for instance transcript variant 1 has 6 transcript exons and 4 coding exons, the strand (-), genetic associations such as autism and bladder cancer, and drugs that interact with the gene such as Cisplatin can all found here. Interestingly we only see 4 transcripts under UCSC genes and we see 8 under Ensembl. However, as we see in exploring this gene in option b. below at the NCBI Genome Reference Consortium, it shows 11 Ensembl transcripts for this gene which is exactly what the Ensembl database says.
         2. Another interesting point which may or may not be a bug is when we view the Short Genetic Variants by setting [All SNPs(151)](https://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=818179607_rLwHqaVTIRbJZDrtfHFB82mla5mo&c=chr11&g=snp151) to full and then selecting the variant rs# [**rs104894228**](https://genome.ucsc.edu/cgi-bin/hgc?hgsid=818174999_fMu2exunJ89SYjUA8e1bovaUiZED&c=chr11&l=534280&r=534291&o=534285&t=534286&g=snp151&i=rs104894228) it will take us to this [link](https://genome.ucsc.edu/cgi-bin/hgc?hgsid=818179607_rLwHqaVTIRbJZDrtfHFB82mla5mo&c=chr11&l=534281&r=534292&o=534285&t=534286&g=snp151&i=rs104894228) where it thinks the reference allele is a G but in fact this is incorrect because if you click further on [dbSNP rs104894228](https://www.ncbi.nlm.nih.gov/snp/rs104894228) it takes you to NIH’s website which correctly attributes the C>A / C>G / C>T variants and under the ‘Change’ column we correctly see the HGVS representation NC\_000011.9:g.534286C>A as well and the newest version NC\_000011.10:g.534286C>A. UCSC is most likely listing the coding strand reference. We can also review the Clinical Significance for each allele variant. Number 1 being Costello syndrome. For this specific variant we can see various connections to certain types of cancer such as Multiple myeloma, Neoplasm of the breast, AML, CLL, and Pancreatic adenocarcinoma.
         3. I also was interested in looking up how I could obtain a large list of Reference SNP id numbers (rs #) given a large list of in the format:

Accession.version:{type of reference}[Start position in 1 indexing]{Ref>Alt} as in below:

NC\_000011.9:g.534286C>A

NC\_000011.9:g.534287GC>AA

NC\_000011.9:g.534289G>C

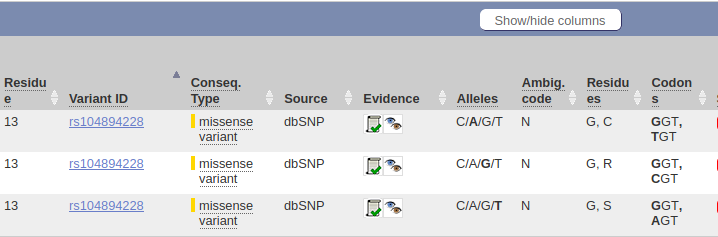
…

I found a resource for batch database request from Heng Li [here](https://www.biostars.org/p/8920/) where he is very adamant about using the `bin` field when querying the UCSC Genome Browser. I wrote a bash script that would take these HGVS inputs and call mySQL queries on each one to obtain the rs #s and write out a tab delimited file with the results where the first column is the HGVS column. The script uses Heng Li’s perl script to query. My script can be found [here](https://github.com/BJWiley233/Practical-Computer-Concepts-Files/blob/master/linux_shell/variant/get_rs_ids.sh) (it’s definitely a hack because of delimiter issue) and the example input file with the list of HGVS representations [here](https://github.com/BJWiley233/Practical-Computer-Concepts-Files/blob/master/linux_shell/variant/snps.txt). This only works for Accessions with the format prefix NC\_0000 obviously. With the results we can see that this specific variant rs id is **rs104894228** just as using the interface.

EDIT: I found out this doesn’t work entirely because Heng’s script does not allow for adding a WHERE clause to filter on the reference and alternate bases, and I do not know how to hack Perl code. Especially his. My bash script `get\_rs\_ids.sh` will pull rs\_ids based on that position but not to the particular reference and alternate bases. I found this out by adding random generated variants to end of the list as well as I variant I manually found and altered the reference base (second variant in list) and it returned rs\_ids incorrectly. The R script that I added uses the ‘myvariant’ package to obtain the rs\_ids and I used the Python function source code from [here](https://bamnostic.readthedocs.io/en/latest/bamnostic.html#bamnostic.bai.reg2bins) to write an Rcpp function I created in file `binFromRange.cpp` to obtain the bins to query UCSC for peptide information.

* 1. A second interface option is to visit NCBI’s Genome Reference Consortium webpage @ <https://www.ncbi.nlm.nih.gov/grc>
     1. Navigate to Human > Scroll down and choose GRCh37.p13 tab > Sort by Chr > Page to first region for Chromosome 11 (Page 5) and click any link under column ‘Region Name’
     2. This brings you to a visualization like the UCSC Genome Browser. Next to find our position we can insert the position 534286 next to ‘find’
     3. When we scroll down we can see two sections right away. We see Genes annotated by NCBI and also by Ensembl. There are a lot of genes when zoomed out at 20kb intervals but if we look where the lines flow down from our position we can see “HRAS[+8]” under NCBI genes. If we click there we see, similarly to UCSC that there are 4 transcripts and also 4 isoform transcripts that potentially code for different Ras proteins. By hovering over one of the transcripts on the HRAS title for the transcripts we can then click “View GeneID”: [3265 (HRAS)](https://www.ncbi.nlm.nih.gov/gene/3265) which brings us to NCBI’s Gene database entry for HRAS. Here we see a lot of information such as that from UCSC Genome Browser indicating this gene is a member of the Ras oncogene family. The gene is highly expressed in the brain and skin. There are a lot of resources under Bibliography. Most of the titles are with respect to the HRAS role in tumorigenesis (breast, epithelial-myoepithetlial carcinoma, bladder) as well as signaling pathways (MEK-ERK, MAPK and Notch). An interesting article is number 7 on PubMed’s website about the mutation [p.Gly13Cys](https://onlinelibrary.wiley.com/doi/abs/10.1002/ajmg.a.38110) for this variant which I confirm programmatically below in section c. This residue mutation in the paper is hypothesized to be the cause for both boys having Costello syndrome.
     4. One thing to mention summarized in findings in R using Biomart is that back under the Consortium page under position 534276 we can see when following the position lines down under Ensembl Genes and by expanding ENSG00000174775 we see there are 11 transcripts with prefix ENST and 6 transcripts with prefix ENSP. I am not sure what the prefixes for ENSP represent but when clicking any of the links we only see the 11 transcripts in the transcript table [here](https://uswest.ensembl.org/Homo_sapiens/Transcript/Summary?g=ENSG00000174775;r=11:532242-535538;t=ENST00000493230). Only 6 of the transcripts code for protein variation. All of the proteins are relatively the same length between 170 and 190 amino acids. They have various amounts of exons 5-7 but only 4 coding exons each similar to the other Ras family genes (KRas and NRas). By clicking any of the links under the Protein Column such as this one [here](https://uswest.ensembl.org/Homo_sapiens/Transcript/ProteinSummary?db=core;g=ENSG00000174775;r=11:532242-535538;t=ENST00000417302), we can see the protein domains. This is not the best view as we can’t really see all the domains on this particular transcript. It really is just for references on the GTPase family. It is easier to just perform a Google search and review some publications. I found HRas has a P-loop NTPase, Small GTP binding domain, Small GTPase, and Small GTPase Ras-type domain at the [AGCOH](http://atlasgeneticsoncology.org/Genes/HRASID108.html) obtained from the InterPro database at EMBL Hinxton, a contender for my PhD applications.
  2. A third way to obtain information about this variant is programmatically. I chose R. In the R script [here](https://github.com/BJWiley233/Practical-Computer-Concepts-Files/blob/master/linux_shell/variant/HGVS.R) I read in the same text document with multiple variant representations where the first one is the one in question from your e-mail. I used the library “myvariant” to obtain the rs\_ids and also queried Ensembl’s SNP database using the library “biomaRt” to see if we could get the associated protein variant. Unfortunately, when querying the rs\_ids from Ensembl we get a line for each of the protein variants for all alternate allele possibilities and the specific peptide variant is not decipherable (see ensembl\_pep[1:10, ] variable in R script which also prints when running) for a particular HGVS variant. So, I only used biomaRt to obtain the Pubmed ids and links for those, the phenotype\_descriptions for related phenotype and alternate Ensembl transcript ids; we see 8 of them (16 are listed as the Ensembl ID changed for the GRCh38 assembly). Interestingly the phenotype\_descriptions from Ensembl include a fair number of skin neoplasms which coincides that HRas is express highly in epithelial cells.
     1. I found a creative way to query the variant peptide coding alleles using the complement of the reference and alternates allele from the HGVS variant for NC\_000011.9:g.534286C>A. You can see it in the loop “for(row in 1:nrow(data))” in my R script which steps through rows with rs\_ids and queries snpCodingDb from UCSC Genome browser using RMySQL. Here we find the C>A variant which according to the interface [here](https://genome.ucsc.edu/cgi-bin/hgc?hgsid=818693121_Y301Oc3Xabnoats4N3NIZnAzp8aD&c=chr11&l=534285&r=534286&o=534285&t=534286&g=snp151&i=rs104894228) is on the negative (-) strand. The coding strand for the mRNA is then a G>T variant. The trick is as follows:
        1. Since the alleles, codons, and peptides are in order in the columns queried from the UCSC Database we can match the allele indices with the mRNA from the variants (Should only be 2 indices per variant for SNPs). Then using those indices only keep those indices in the columns alleles, codons, and peptides. I confirmed this works by obtaining [sample variants](https://raw.githubusercontent.com/biocommons/hgvs/master/tests/data/gcp/ADRA2B-dbSNP.tsv) from biocommons ‘rgvs’ package. The column ‘Peptide2’ from my HGVS\_protein.csv matches all the HGVSp column from biocommons sample except the biocommons test uses ‘Ter’ for the stop codons where the R library ‘seqinr’ uses ‘Stp’ for stop codons. This could easily be manipulated in R script.
     2. Finally we find from the R script’s first output indicated in the file HGVS\_protein.csv that this specific variant from your email is a Glycine(GGT) to Cysteine(TGT) residue mutation in the HRas protein. I used the ‘seqinr’ library to convert the 3 letter amino acid code from 1 letter. We can confirm this on the two of the interface options above from UCSC and Ensembl databases:
        1. [UCSC](https://genome.ucsc.edu/cgi-bin/hgc?hgsid=819080085_S8uao5cxmbGXnE5nta4wAOrtDCXi&c=chr11&l=534280&r=534291&o=534285&t=534286&g=snp142&i=rs104894228): HRAS (NM\_001130442): [missense\_variant](http://sequenceontology.org/browser/current_release/term/SO:0001583) G (**G**GT) --> C (**T**GT)
        2. [Ensembl](https://uswest.ensembl.org/Homo_sapiens/Transcript/ProtVariations?db=core;g=ENSG00000174775;r=11:532242-535576;source=dbSNP;t=ENST00000311189;v=rs104894228;vdb=variation;vf=90533015): We can page back from the image below after confirming from the image the residue variants [here](https://uswest.ensembl.org/Homo_sapiens/Variation/Explore?db=core;r=11:533786-534786;v=rs104894228;vdb=variation;vf=90533015) under Variant allele A [ENSP00000309845](https://uswest.ensembl.org/Homo_sapiens/Transcript/ProtVariations?db=core;source=dbSNP;t=ENSP00000309845.7;v=rs104894228;vdb=variation;vf=90533015).7:p.Gly13Cys that is indeed the residue variant from the PubMed publication on the two boys with Costello syndrome.

**rs104894228**



Files are all available in [GitHub folder](https://github.com/BJWiley233/Practical-Computer-Concepts-Files/tree/master/linux_shell/variant):

Script Files:

* get\_ebi\_assemblies.sh (Input: list of HGVS variants, Output: sql query file)
* get\_rs\_ids.sh (Input: list of HGVS variants [snps.txt](https://github.com/BJWiley233/Practical-Computer-Concepts-Files/blob/master/linux_shell/variant/snps.txt), Output: rs\_ids.tsv. Downloads Heng Li’s [batch](https://github.com/lh3/misc/blob/master/biodb/batchUCSC.pl) perl script)
* [HGVS.R](https://github.com/BJWiley233/Practical-Computer-Concepts-Files/blob/master/linux_shell/variant/HGVS.R) (No input. Gets list of HGVs variants in [snps.txt](https://github.com/BJWiley233/Practical-Computer-Concepts-Files/blob/master/linux_shell/variant/snps.txt) from my Github account. First one being variant from e-mail. Requires [binFromRange.cpp](https://github.com/BJWiley233/Practical-Computer-Concepts-Files/blob/master/linux_shell/variant/binFromRange.cpp). Outputs HGVS\_protein.csv, HGVS\_ensembl\_pubmed.csv, HGVS\_ensembl\_pheno.csv, HGVS\_ensembl\_transc.csv)
  + - Wall time for 109 HGVS variants:

real 0m28.613s

user 0m15.713s

sys 0m1.160s

* binFromRange.cpp (Takes reg2bins from biocommons and makes a it C++ function)

Input Text files:

* snps.txt

Output TSV and CSV files:

* rs\_ids.tsv
* HGVS\_protein.csv (All HGVS with NC\_000011.9:g.534286C>A first)
* HGVS\_ensembl\_pubmed.csv (Only for NC\_000011.9:g.534286C>A)
* HGVS\_ensembl\_pheno.csv (Only for NC\_000011.9:g.534286C>A)
* HGVS\_ensembl\_transc.csv (Only for NC\_000011.9:g.534286C>A)

This Word/odt file.

References:

1. UCSC Genome Browser: Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006.
2. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11.
3. Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H. C., Agarwala, R., McLaren, W. M., Ritchie, G. R., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., Matthews, L., … Hubbard, T. (2011). Modernizing reference genome assemblies. PLoS biology, 9(7), e1001091. <https://doi.org/10.1371/journal.pbio.1001091>
4. Fiona Cunningham, Premanand Achuthan, Wasiu Akanni, James Allen, M Ridwan Amode, Irina M Armean, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, Carla Cummins, Claire Davidson, Kamalkumar Jayantilal Dodiya, Astrid Gall, Carlos García Girón, Laurent Gil, Tiago Grego, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G Izuogu, Sophie H Janacek, Thomas Juettemann, Mike Kay, Matthew R Laird, Ilias Lavidas, Zhicheng Liu, Jane E Loveland, José C Marugán, Thomas Maurel, Aoife C McMahon, Benjamin Moore, Joannella Morales, Jonathan M Mudge, Michael Nuhn, Denye Ogeh, Anne Parker, Andrew Parton, Mateus Patricio, Ahamed Imran Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Eloise Stapleton, Marek Szuba, Kieron Taylor, Glen Threadgold, Anja Thormann, Alessandro Vullo, Brandon Walts, Andrea Winterbottom, Amonida Zadissa, Marc Chakiachvili, Adam Frankish, Sarah E Hunt, Myrto Kostadima, Nick Langridge, Fergal J Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Daniel M Staines, Stephen J Trevanion, Bronwen L Aken, Andrew D Yates, Daniel R Zerbino, Paul Flicek, Ensembl 2019, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D745–D751, <https://doi.org/10.1093/nar/gky1113>
5. Jean-Loup Huret, Mohammad Ahmad, Mélanie Arsaban, Alain Bernheim, Jérémy Cigna, François Desangles, Jean-Christophe Guignard, Marie-Christine Jacquemot-Perbal, Maureen Labarussias, Vanessa Leberre, Anne Malo, Catherine Morel-Pair, Hossein Mossafa, Jean-Claude Potier, Guillaume Texier, Franck Viguié, Sylvie Yau Chun Wan-Senon, Alain Zasadzinski, Philippe Dessen, Atlas of Genetics and Cytogenetics in Oncology and Haematology in 2013, Nucleic Acids Research, Volume 41, Issue D1, 1 January 2013, Pages D920–D924, <https://doi.org/10.1093/nar/gks1082>
6. Wang, M., Callenberg, K. M., Dalgleish, R., Fedtsov, A., Fox, N. K., Freeman, P. J., Jacobs, K. B., Kaleta, P., McMurry, A. J., Prlić, A., Rajaraman, V., & Hart, R. K. (2018). hgvs: A Python package for manipulating sequence variants using HGVS nomenclature: 2018 Update. Human mutation, 39(12), 1803–1813. <https://doi.org/10.1002/humu.23615>
7. Parnell LD, Lindenbaum P, Shameer K, Dall'Olio GM, Swan DC, Jensen LJ, et al. (2011) BioStar: An Online Question & Answer Resource for the Bioinformatics Community. PLoS Comput Biol 7(10): e1002216. https://doi.org/10.1371/journal.pcbi.1002216