Variant Effect Prediction Training Course

Johor (Malaysia), August 27-30, 2018

Describing variants - HGVS nomenclature

The standard to describe, and store in databases, variants found in DNA sequences are the "*HGVS recommendation for the description of sequence variants*" (den Dunnen et al. (2016). Hum Mutat 37:564). On http://varnomen.HGVS.org you can find a very detailed description of the HGVS nomenclature and a free copy of the paper.

As background of the recommendations you can use the slides from the introduction lecture, the "HGVS Simple" page on the HGVS website (under Tab Background Materials > HGVS simple) or the presentation available on the website (Background Materials > Educational Material > Slide presentation).

For all **ToTry's** note that you can query the resource using any variant you are interested in. The examples we give is just for those that lack inspiration. We encourage participants to check variants from their own work.

Support tools

There are several tools available to check or generate HGVS descriptions and/or to go from a description based on one reference sequence to that based on another reference sequence. In this task we will use the Mutalyzer tools (wwwMutalyzer.nl) but there are alternatives like e.g. the VariantValidator (variantvalidator.org/), ClinGen Allele Registry (reg.clinicalgenome.org) and others. The tools usually have options to test variants one by one using the website or in a "Batch Mode", facilitating the analysis of thousands of variants at the same time.

While our **ToTry's** use Mutalyzer, we suggest to try another tool in parallel. Ultimately, use the one you like best, i.e. easiest to work with to get the results you need!

Below a series of tasks we prepared. Time available will be too short to complete all. It is therefore probably wise to make sure to try a few of each or to quickly scan through and select the task which is most relevant for your work.

ToTry's

- checking variant descriptions (Mutalyzer Name Checker)
- generating variant descriptions (Mutalyzer Variant Description Extractor)
- to another reference sequence using (Mutalyzer Position Converter)
- SNP details using (Mutalyzer SNP Converter)
- make HGVS variant descriptions (based on practical examples)

Checking variant descriptions using Mutalyzer's Name Checker

The Name Checker tool can be used to check whether or not a description is correct, i.e. follows the recommendations. This tool can be handy when you doubt whether a description is correct or to check descriptions you have generated.

ToTry: go to the Mutalyzer site (www.Mutalyzer.nl) and try the "Name Checker" (top of your screen) to check a variant description.

- enter the description **LRG_123:g.12409G>C** in the query field and click the blue button "Check variant description".
After a few seconds, the result will indicate the description is not correct, the reference sequence (LRG_123) does not contain a G at position 12409 (but an A).

- repeat the query using **LRG_123:g.12409A>C**.
The result will show the description is correct and falls in the annotated coding DNA reference sequence for transcript 1 of the UNC93B1 gene (LRG_123t1:c.976A>C), with p.(Asn326His) as the predicted missense change at the protein level, and creating restriction enzyme recognition sites for HaeIII, NlaIV, and Sau96I.

- repeat the query using **NM_004006.2:c.5209C>T**.
The result will show the description is correct and will give the reference and predicted variant protein sequence.

- repeat the query using **NC_000011.9:g.111959693T>G**.
The result will show the description is correct, show the genomic reference sequence around the variant, whether it affects transcripts/proteins, etc.

- repeat the query using **NM_004006.2:c.3276+2T>A**.
The result gives an error; do you understand why the description is not correct?
Next try **NC_000023.10(NM_004006.2):c.3276+2T>A.**
You now gave a proper reference sequence and the description was approved.

- repeat the query using **NC_000023.10(NM_004006.2):c.3276+3del**.
The correct description is NC_000023.10(NM_004006.2):c.3276+3del, although the warning given is confusing and the correct description only returned at the genomic level (g.).

NOTE: when Mutalyzer is slow you can try the test-server (test.Mutalyzer.nl) but please note this tool is under development and may give different results compared to the official Mutalyzer release.

- from the Menu, go to *"Batch Jobs"*, select the "*Name Checker"*. and check the file format to use (lower right). Make a file containing a few variants and give it a try.

Generating variant descriptions using Mutalyzer's Variant Description Extractor

The Variant Description Extractor tool generates HGVS description by comparing a Reference and a Sample sequence.

ToTry: go to the Mutalyzer website > select the "Description Extractor"

- take a sequence (the **Test sequence** below or one of your own interest)

Test sequence

```
ATGCTTTGGTGGGAAGAAGTAGAGGACTGTTATGAAAGAGAAGATGTTCAAAAGAAAACA
TTCACAAAATGGGTAAATGCACAATTTTCTAAGTTTGGGAAGCAGCATATTGAGAACCTC
TTCAGTGACCTACAGGATGGGAGGCGCCTCCTAGACCTCCTCGAAGGCCTGACAGGGCAA
```

- add the sequence in to the "Reference sequence" pane and in to the "Sample Sequence" pane

- make a change and click the blue "Extract variant description"
Is the result as expected?, do you understand the format of the variant description?

- other changes to try: change one nucleotide to another nucleotide, remove one or more nucleotides, duplicate a few nucleotides, insert a few nucleotides, etc.
NOTE: either reset the sample sequence before you add a next change or accept that you get the combined result of several changes.

- *optional* (for those interested): where is this sequence from?

To another reference sequence using Mutalyzer's Position Converter

The Position Converter can be used to get from one reference sequence to another, e.g. from c. to g. or vice versa. Note that *not all reference types are supported* and also that the tool *does not check whether the description of the variant given is correct* or not. The tool blindly assume that what you enter is correct.

ToTry: go to the Mutalyzer website > select the "Position Converter"

- under "Build", select the reference sequence to use ("Homo sapiens - GRCh37/hg19")

- enter the description "**NM_030930.2:c.976A>C**" and click the blue button labeled "Convert variant description".
The result (NC_000011.9:g.67764185T>G) shows that, in genome build hg19, the variant is located on chromosome 11 at position g.67764185 in the UNC93B1 gene. Since the variant is described as T>G, while A>C was given on coding DNA level, it can be concluded the gene is located on the minus strand (in antisense orientation).

- go back to the Position Converter start page and change the genome build to Homo sapiens - GRCh38/hg38") and enter again "**NM_030930.2:c.976A>C**" .
The result (NC_000011.10:g.67996715T>G) shows that, in genome build 38, the variant is located on chromosome 11 at position g.67996715. Note the version number of the reference sequence (NC_000011) changed from .9 to .10.

- try **NM_004006.2:c.5209C>T**.
The result (NC_000023.11:g.32362904G>A) shows that, in the genome build 38, the variant is located on the X-chromosome at position g.32362904, in the DMD gene. Since from this gene several RNA transcripts are generated, the "Position Converter" give a series of "c." variant descriptions.

- to "see" variant NM_004006.2:c.5209C>T in relation to the encoded protein sequence go to the LOVD database for DMD (http://www.LOVD.nl/DMD), select the option "Refseq URL > Genomic reference sequence" and find nucleotide c.5209. What is the predicted effect of the C>T variant on protein translation?
The predicted effect is p.(Gln1737).*
NOTE: to "see" the variant you can also enter NM_004006.2:c.5209C>T directly in the UCSC genome browser (http://genome-euro.ucsc.edu, select build hg19). When the LOVD track is active (activate under *"Phenotype and Literature"*) you can also see whether the variant has been reported in an LOVD database.

- take the variant **NM_004006.2:c.5209C>T**, set the genome build to "Homo sapiens GRCh37/hg19" and convert.
The result (NC_000023.10:g.32381021G>A) indicates that on this older genome build the variant is located at position g.32381021. Realize how important it is when using variant descriptions that you know the reference sequence used (here genome build) as well as the reference sequence/genome build used by the database you may want to query.

NOTE: the "g." genome build conversion described above using Mutalyzer assumes

the “c.” description did not change. When the variant is ***in or close to*** an exon this assumption is usually correct. However, when far from a gene or deep inside a large intron chances increase this assumption is not correct. For such cases it is better to use specific tools to get from one genome build to another, e.g. the UCSC LiftOver tool (http://genome-euro.ucsc.edu/cgi-bin/hgLiftOver).

- try **chr11:g.111959693G>T**
When a genome build is selected Mutaylzer accepts “chr11” as reference sequence since it is able to derive the NC_ reference sequence.
The result shows this variant may affect transcripts from 3 different genes.

- based on the variant descriptions can you tell, in relation to the protein translation, where these variants are located?
Relative to C11ORF57 variants are in the 3'UTR or downstream of the gene (c. descriptions), relative to SDHD in the coding region or intronic (c. descriptions) and relative to TIMM8B variants are in the 5'UTR or upstream of the gene (c.- descriptions)*

- try **chr11:g.111959693G>T** in the name Checker. Do you understand the error?
The Name Checker requires a reference sequence, when “chr11” is used the tool does not know the genome build you want to use.

- from the Menu, go to *“Batch Jobs”,* select the *“Position Converter”* and check the file format to use (lower right). Make a file containing a few variants and give it a try.

SNP details using Mutalyzer's SNP Converter

The SNP converter can be used to get from a dbSNP identifier (rs number) to descriptions of the variant based on genomic (g.) and coding DNA (c.) reference sequences. Note that the same information can be obtained using the dbSNP database (www.ncbi.nlm.nih.gov/SNP/).

ToTry: go to the Mutalyzer site and try the "SNP Converter" to get from rs IDs to variant descriptions.check a variant description.

- where (which chromosome, which gene) is the variant **rs72468685** located?
*The variant is on the X-chromosome (reference sequence starts with NC_0000**23**) in an intron of the DMD gene (NM_004006.2:c.2293-182A>G). The name of the gene is not displayed but can be retrieved using variant description in the Name Checker or Position Converter.*

- from the Menu, go to *"Batch Jobs",* select the *"SNP Converter"* and check the file format to use (lower right). Make a file containing a few variants and give it a try.

Make HGVS variant descriptions

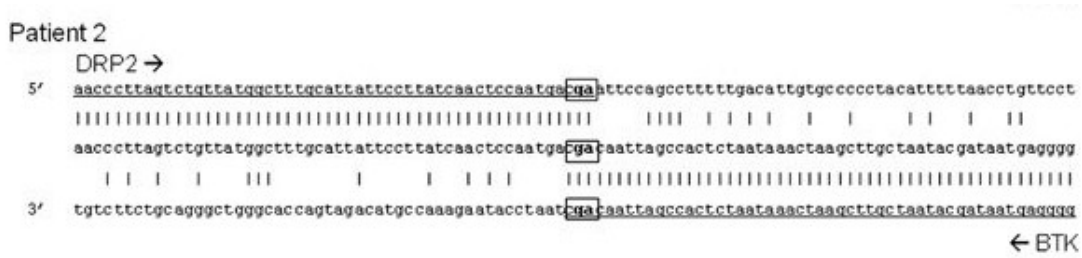
For the expert we have selected some examples of variants detected in practice. The question for all will be to generate a correct HGVS description of the variant. Helpfull tools include;

- UCSC's Blat (genome-euro.ucsc.edu/cgi-bin/hgBlat)
  copy/paste or re-type the break point sequence to get genomic coordinates (*select the correct genome build!*)
- Mutalyzer's Name Checker (www.mutalyzer.nl/name-checker)
  describe the deletion and check whether it is correct. Mutlayzer will check proper application of the 3'rule and display the break point sequence (compare with the original).
- Mutalyzer's Position Converter (www.mutalyzer.nl/position-converter)
  get from a "g." description (genomic nucleotide positions) to a "c." descriptions and the gene's affected by the variant.

ToTry

- a colleague asks which reference sequence to use and how to correctly describe the nucleotide and amino acid variant for IDH1 p.R132C. What do you suggest?
Without the reference sequence you can only guess. Luckily the amino acid change can be caused by one substitution only, so it is most likely a substitution. Based on the data available from the IDH1 gene variant database (www.LOVD.nl/CDH1) the reference sequence might be NM_005896.2. Indeed it codes for an "Arg (R)" residue at position 132. Variant NM_005896.2:c.394C>T gives the predicted protein consequence p.(Arg132Cys).
***NOTE:** using IDH1:Arg132Cys in the UCSC browser you can cehck whether this description is unique or not.*

- in patient 2 (from Arai et al., 2011 J Hum Genet. 56:577) a large deletion involving several genes was detected. Using FISH, arrays, PCR and sequencing the following break point sequence could be defined.



What is the HGVS description of this variant?

- in patient 3 (from Arai et al., 2011 J Hum Genet. 56:577) a large deletion involving several genes was detected. Using FISH, arrays, PCR and sequencing the following break point sequence could be defined.
NOTE: a difficult case, the Figure contains an error.

What is wrong with the Figure?
What is the HGVS description of this variant?

Johan den Dunnen, August 2018