# HGVS Basics

Steven Harrison

Biocurator Call

Feb 14 2019

# http://varnomen.hgvs.org/

## Current Recommendations

| | | |
|---|---|---|
| General | DNA | RNA |
| Protein | Uncertain | Checklist |
| Open Issues | | |

## Background Material

| | | |
|---|---|---|
| Basics | Reference Sequences | Standards |
| Numbering | Community Consultation | HGVS Simple |
| Educational Material | Glossary | |

# Basics – REFERENCE SEQUENCE

In order to have a unique description (that prevents confusion), you must include a reference sequence when using HGVS

*BRCA1* c.4366A>G

| | Option 1 | Option 2 |
|---|---|---|
| NM_007294.3 | **c.4366A>G** | c.4358-2777A>G |
| NM_007300.3 | c.4429A>G | **c.4366A>G** |
| Genomic (GRCh37) | chr17:g.41228623T>C | chr17:g.41231408T>C |

Reference sequence used **must contain** the variant residue described - a coding DNA reference sequence does not contain intron and therefore **cannot be used** to describe intron variants

not correct: NM_004006.2:c.357+1G>A

correct: NG_012232.1(NM_004006.2):c.357+1G>A

**NM_000363.4(TNNI3):c.373-10G=**

Variation ID: 36881
Review status: ★★☆☆ criteria provided, multiple submitters, no conflicts

## Interpretation

Go to:

Clinical significance: Benign/Likely benign
Last evaluated: Jun 14, 2016
Number of submission(s): 6
Condition(s):
- Primary familial hypertrophic cardiomyopathy [MedGen - Orphanet - Orphanet - OMIM]
- Ciliary dyskinesia [MedGen - Orphanet - OMIM - Human Phenotype Ontology]
- Nemaline Myopathy, Recessive [MedGen]

See supporting ClinVar records

## Allele(s)

Go to:

**NM_000363.4(TNNI3):c.373-10G=**

Allele ID: 45542
Variant type: single nucleotide variant
Cytogenetic location: 19q13.4
Genomic location:
- Chr19: 55154216 (on Assembly GRCh38)
- Chr19: 55665584 (on Assembly GRCh37)

HGVS:
- NG_007866.2:g.8517G=
- NG_011829.2:g.23G=
- NM_000363.4:c.373-10G=
- NC_000019.10:g.55154216C= (GRCh38)
- LRG_432t1:c.373-10G=
- NC_000019.9:g.55665584A>C (GRCh37)
- NG_007866.2:g.8517=
- NM_000363.4:c.373-10T>G
- LRG_432:g.8517G=

ClinVar asks for genome build (even if submitting all variants with NM RefSeq HGVS) because of this issue with intronic variants

# Basics – REFERENCE SEQUENCE

- Only public files from NCBI or EBI are accepted as reference sequence files
  - Approved reference sequence formats include:
    - NC_# (e.g. NC_000023.10)
    - LRG_# (e.g. LRG_199, LRG_199t1)
    - NG_# (e.g. NG_012232.1)
    - NM_# (e.g. NM_004006.2)
    - NR_# (e.g. NR_002196.1)
    - NP_# (e.g. NP_003997.1)

Type of reference sequence indicated by letter used:

**c.** (coding); **g.** (genomic); **m.** (mitochondrial); **n.** (non-coding);

**r.** (RNA); **p.** (protein)

# Basics – General Information

- All variants should be described at **the DNA level**. Descriptions at the RNA and/or protein level may be given in addition
- Descriptions should make clear whether the change was **experimentally determined** or **theoretically deduced** by giving predicted consequences in parentheses

  - **NP_003997.1:p.(Trp24Cys)** means amino acid Trp24is predicted to change to a Cys (no experimental proof, e.g. based on DNA level data)
  - **NP_003997.1:p.Trp24Cys** means amino acid Trp24 is changed to a Cys (confirmed via RNA or protein sequence analyzed)

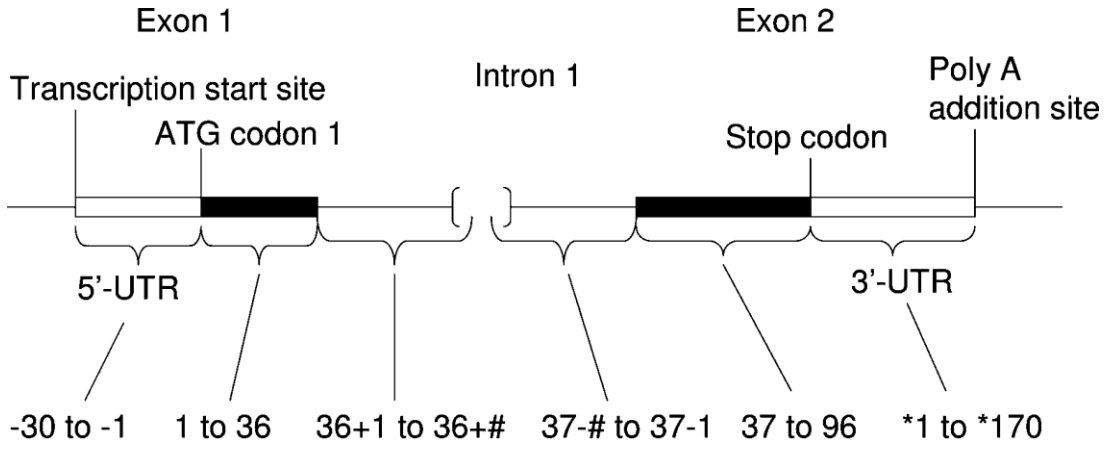
# Basics – General Information

- **Prioritization**: when a description is possible according to several types, the preferred description is: (1) deletion, (2) inversion, (3) duplication, (4) conversion, (5) insertion

- Descriptions at DNA, RNA and protein level differ:
  - **DNA-level** 123456A>T: number(s) referring to the nucleotide(s) affected, nucleotides in CAPITALS
  - **RNA-level** 76a>u: number(s) referring to the nucleotide(s) affected, nucleotides in lower case
  - **protein level** Lys76Asn: the amino acid(s) affected in 3- or 1-letter followed by a number (* **three-letter** amino acid code is preferred)

# Variant nomenclature: cDNA

- Nucleotide 1 is the A of the ATG initiation codon (there is no c.0)
- The nucleotide 5’ of the ATG initiation codon is -1, the previous -2, etc.
- The nucleotide 3’ of the stop codon is *1, the next *2, etc.
- Intronic nucleotides
  - 5’ end of the intron: the number of last coding nucleotide of the preceding exon, a plus sign and the position within in the intron, e.g., c.36+1G, c.36+2T
  - 3’ end of the intron: the number of the first coding nucleotide of the following exon, a minus sign and the position upstream in the intron, e.g. c.37-1G, c.37-2A




# indicates any positive integer number

Ogino et al. (2007), *JMD*

# Substitutions

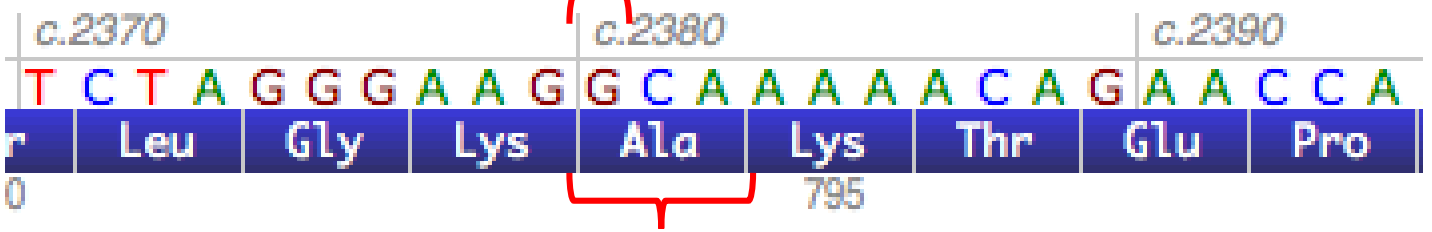
- Substitutions designated by: >

- Examples:
  - Genomic - NC_000011.10:g.47342698G>A
  - cDNA - NM_000256.3:c.1504C>T
  - RNA - NM_000256.3:r.1504c>u
  - Protein - NP_000247.2:p.(Arg502Trp)

# Deletions

- Format: **“prefix”“position(s)_deleted”“del”**, e.g. g.123_127del

c.2380delG or c.2380del




c.2380_2382delGCA or c.2380_2382del

NOT c.2380del3

- For all descriptions the **most 3’ position** possible of the reference sequence is arbitrarily assigned to have been changed (**3’rule**)

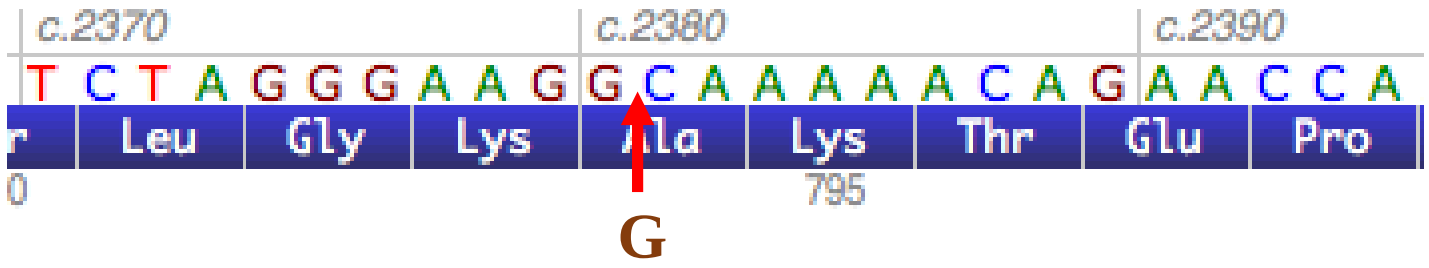
# Large Deletions


- If exons 30 and 31 are deleted and exons 29 and 32 are known to NOT be deleted, deletion is named:
  - c.(3190+1_3191-1)_(3490+1_3491-1)del

- If exons 30 and 31 are deleted and status of surrounding exons is unknown, deletion is named:
  - c.(?_3191-1)_(3490+1_?)del

# Duplications

- Format: **“prefix”“position(s)_duplicated”“dup”** e.g. g.123_345dup

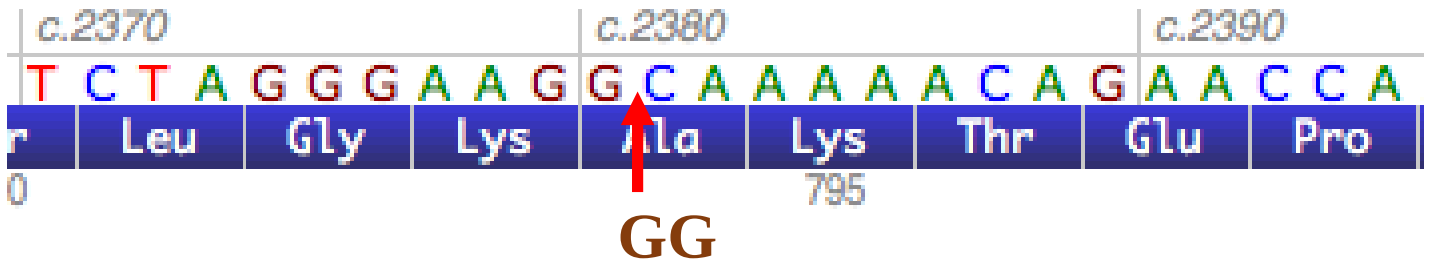



c.2380dupG or c.2380dup
**NOT** c.2380_2381insG

- For all descriptions the **most 3’ position** possible of the reference sequence is arbitrarily assigned to have been changed (**3’rule**)

# Duplications

- Format: **“prefix”“position(s)_duplicated”“dup”** e.g. g.123_345dup




c.2379_2380dupGG or c.2379_2380dup

**NOT** c.2380_2381insGG

- For all descriptions the **most 3’ position** possible of the reference sequence is arbitrarily assigned to have been changed (**3’rule**)

# Large Duplications

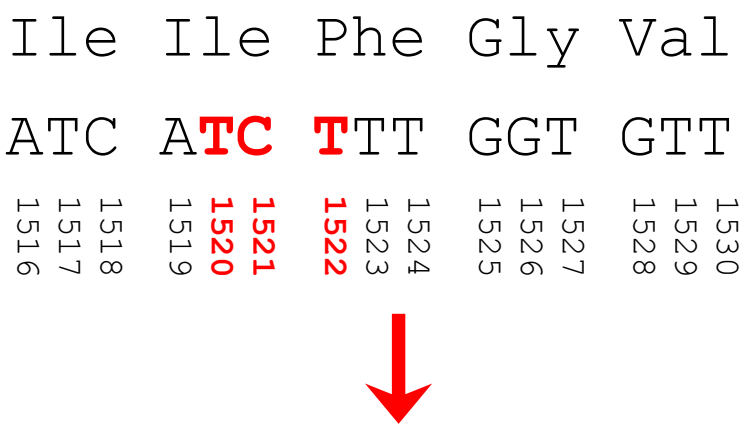

- If exons 30 and 31 are duplicated and exons 29 and 32 are known to NOT be duplicated, variant is named:
  - c.(3190+1_3191-1)_(3490+1_3491-1)dup

- If exons 30 and 31 are duplicated and status of surrounding exons is unknown, variant is named:
  - c.(?_3191-1)_(3490+1_?)dup

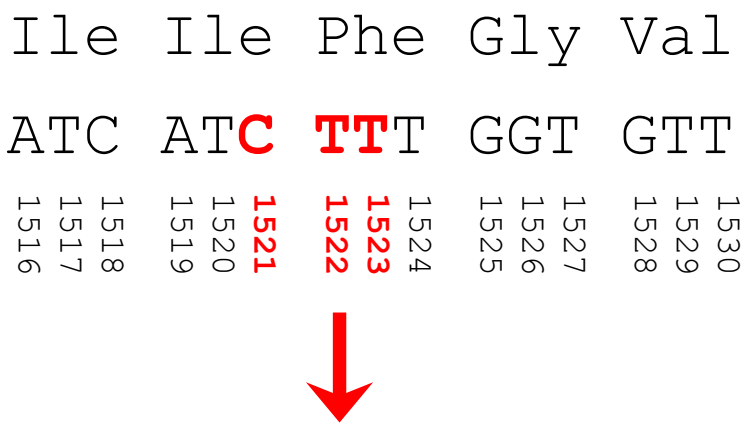
- HGVS notes in this scenario, dup **should** be in tandem

# Which nucleotides are deleted?

**NM_000492.3:c.1520_1522delTCT**




**NM_000492.3:c.1521_1523delCTT**



# Which nucleotides are deleted?

**NM_000492.3:c.1520_1522delTCT**

Ile Ile Phe Gly Val

ATC A**TC** **T**TT GGT GTT

1516 1517 1518 1519 **1520** **1521** **1522** 1523 1524 1525 1526 1527 1528 1529 1530

Ile Ile Gly Val

ATC ATT GGT GTT

**NM_000492.3:c.1521_1523delCTT**

Ile Ile Phe Gly Val

ATC AT**C** **TT**T GGT GTT

1516 1517 1518 1519 1520 **1521** **1522** **1523** 1524 1525 1526 1527 1528 1529 1530

Ile Ile Gly Val

ATC ATT GGT GTT

**Same end result with either deletion**

**Since we don’t know which event occurred, the 3’ most representation is selected**

# Which nucleotides are deleted?

**NM_000492.3:c.1520_1522delTCT**

Ile Ile Phe Gly Val

ATC A**TC** **T**TT GGT GTT

1516 1517 1518 1519 **1520** **1521** **1522** 1523 1524 1525 1526 1527 1528 1529 1530

↓

Ile Ile Gly Val

ATC ATT GGT GTT

**NM_000492.3:c.1521_1523delCTT**

Ile Ile Phe Gly Val

ATC AT**C** **TT**T GGT GTT

1516 1517 1518 1519 1520 **1521** **1522** **1523** 1524 1525 1526 1527 1528 1529 1530

↓

Ile Ile Gly Val

ATC ATT GGT GTT

**Variant: 7-117199644-ATCT-A**

**Genomic databases report with VCF – meaning 5’ most expression**

# Genomic vs cDNA 3’ Rule

- **The “g.” and “c.” variant descriptions differ regarding the deleted nucleotide when applying the 3’ rule (if gene is on the minus strand)**


If you delete one of the G nucleotides (genomic; or C nucleotide at cDNA) ….

# Genomic vs cDNA 3’ Rule

- **The “g.” and “c.” variant descriptions differ regarding the deleted nucleotide when applying the 3’ rule (if gene is on the minus strand)**


If you delete one of the G nucleotides (genomic; or C nucleotide at cDNA) ….

**Genomic HGVS expression would be g.5delG (which corresponds to c.4delC)**

# Genomic vs cDNA 3’ Rule

- **The “g.” and “c.” variant descriptions differ regarding the deleted nucleotide when applying the 3’ rule (if gene is on the minus strand)**

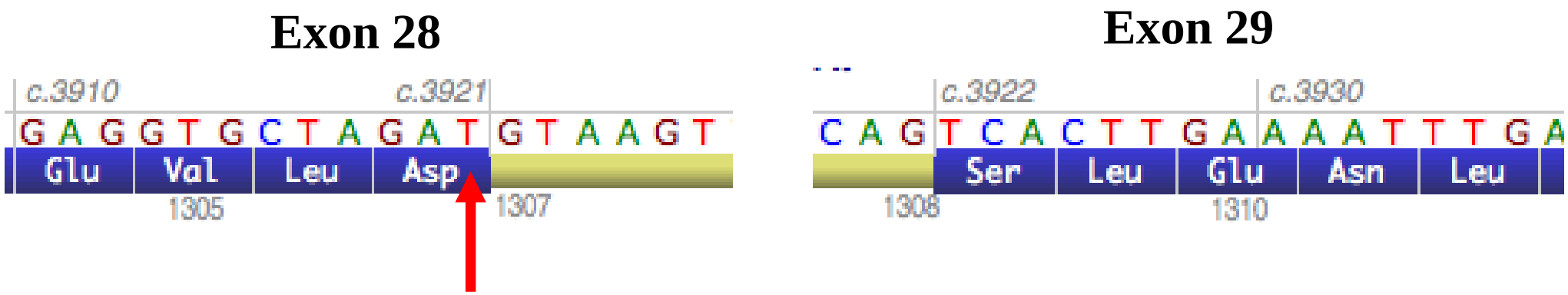

If you delete one of the G nucleotides (genomic; or C nucleotide at cDNA) ….

**Genomic HGVS expression would be g.5delG (which corresponds to c.4delC)**

**cDNA HGVS expression would be c.7delC (which corresponds to g.2delG)**

# 3’ Rule Exception

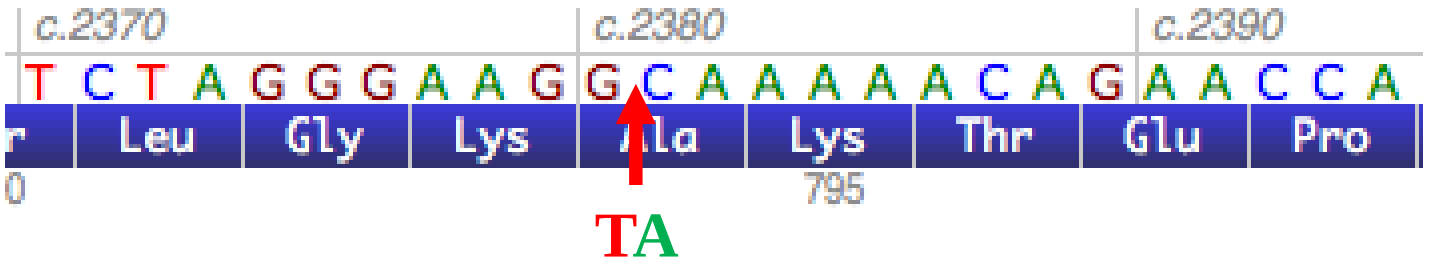
3’ rule is NOT used for deletions around exon/exon junctions when identical nucleotides flank the junction




If you observed a deletion of nucleotide T at c.3921 you would still call the deletion NM_004006.2:c.3921delT **not** c.3922delT (because c.3921 and c.3922 are separated by an intron)

# Insertions

- Format: **“prefix”“positions_flanking”“ins”“inserted_sequence”**, e.g. g.123_124insAGC
- MUST include the nucleotides inserted
- Cannot list only 1 flanking position




c.2380_2381insTA
NOT c.2380_2381ins2
NOT c.2380insTA

# Repeats


- Such changes are described using the format ***"position-first-repeat-unit"“sequence”[number]***" (e.g. c.52CAG[6]) where ***position-first-repeat-unit*** *gives the location of the first unit* of the variable sequence repeat and ***[number]*** *the number of units present* in the allele described.
- This nomenclature uses the 1st repeat NOT the most 3’ repeat

# Alleles

- Alleles indicated by **[ ]** and separated by **;**
- 2 changes, 2 alleles
  - c.[428A>G];[83dupG] (both copies of the gene have a variant)
- 1 allele, several changes (haplotype)
  - c.[12C>G; 428A>G; 983dupG]
- 2 changes, allele status unknown
  - c.428A>G(;)83dupG

# Protein nomenclature

- 3-letter amino acid code is preferred to describe the amino acid residues (Lys vs. K for lysine)
- For all descriptions the ***most C-terminal position possible*** is arbitrarily assigned to have been changed
- Methionine encoded by the translation initiation site (*start codon*) is numbered as residue 1 ("***Met1***" or ***" M1"*** )
- ***"Ter"*** or ***"*"*** designating a translation termination codon (some labs use X)

# Protein nomenclature

- **Silent changes:** p.Leu54Leu or p.=
- **Substitutions:** p.Trp26Cys
- **Nonsense variant:** p.Trp26Ter or p.Trp26*
- **No-stop change:** p.Ter110GlnextTer17 or p.*110Glnext*17
- **In-frame deletions:** p.Gln8del or p.Cys28_Met30del
- **Duplications:** p.Gly4_Gln6dup
- **Insertions:** p.Lys2_Met3insGlnSerLys
- **Frameshifts:** short description: p.Arg97fs
  long description: p.Arg97Profs*23

where the "*Arg97Pro*" describes the substitution of Arg for Pro at position 97, "*fs*" indicating the frameshift and the "**23*" describes the position of the translational termination (stop) codon in the new reading frame (starting with proline as amino acid #1)

# Acknowledgements


- Hana Zouk
- Christina Austin-Tse