Describing variants

"mutation nomenclature"

recommendations for the description of DNA changes

Johan den Dunnen

chair SVD-WG

VarNomen @ HGVS.org

http://www.HGVS.org/varnomen/

HGVS / HVP / HUGO Sequence Variant Description working group

Working Group Members:

- Anne-Francoise Roux (EGT)
- Donna Maglott (NCBI/EBI)
- Jean McGowan-Jordan (ISCN)
- Peter Taschner (LSDBs)
- Raymond Dalgleish (LSDBs)
- Reece Hart (industry)
- Johan den Dunnen (chair)
- HGVS - Marc Greenblatt
- HUGO - Stylianos Antonarakis

Nomenclature

(describing DNA variants)

Stable

Meaningful ***Memorable***

Unequivocal

Definitions

- **prevent confusion**
  - *do not use "mutation"*
    - *use variant, disease-associated variant*
  - *do not use "polymorphism"*
    - *use variant, not disease-associated variant*
  - *do not use "pathogenic"*
    - *use disease-associated, a disease-associated variant*

- **better use neutral terms**
  - *sequence variant*
  - *alteration*
  - *CNV* *(Copy Number Variant)*
  - *SNV* *(Single Nucleotide Variant, not SNP)*

Variant description

the basis ***http:// www.HGVS.org / varnomen***

SPECIAL ARTICLE

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

HGVS Recommendations for the Description of Sequence Variants: 2016 Update

Hum Mutat (2016) 37:564-569

Johan T. den Dunnen,[1*] Raymond Dalgleish,[2] Donna R. Maglott,[3] Reece K. Hart,[4] Marc S. Greenblatt,[5] Jean McGowan-Jordan,[6] Anne-Francoise Roux,[7] Timothy Smith,[8] Stylianos E. Antonarakis,[9] and Peter E.M. Taschner[10] on behalf of the Human Genome Variation Society (HGVS), the Human Variome Project (HVP), and the Human Genome Organisation (HUGO)

HUMAN MUTATION 15:7–12 (2000)

MDI SPECIAL ARTICLE

Mutation Nomenclature Extensions and Suggestions to Describe Complex Mutations: A Discussion

Johan T. den Dunnen[1*] and Stylianos E. Antonarakis[2*]
[1]*MGC-Department of Human and Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands*
[2]*Division of Medical Genetics, University of Geneva Medical School, Geneva, Switzerland*

Consistent gene mutation nomenclature is essential for efficient and accurate reporting, testing, and curation of the growing number of disease mutations and useful polymorphisms being discovered in the human genome. While a codified mutation nomenclature system for simple DNA lesions has now been adopted broadly by the medical genetics community, it is inherently difficult to represent complex mutations in a unified manner. In this article, suggestions are presented for reporting just such complex mutations. Hum Mutat 15:7–12, 2000.

KEY WORDS: complex mutation; mutation detection; mutation database; nomenclature; MDI

THE HUMAN VARIOME PROJECT

HUGO

HGVS HUMAN GENOME VARIATION SOCIETY

THE HUMAN VARIOME PROJECT

HGVS HUMAN GENOME VARIATION SOCIETY

www.HGVS.org/varnomen

Sequence Variant Nomenclature | Recommendations | Background Materials | Recent Additions | Contact Us | Version 15.11

Sequence Variant Nomenclature

Recent Additions

An overview of recent additions, especially those that led to a change of the *HGVS version number*, can be found on the *Versioning page*. The *Open Issues* page shows whether there are proposals open for *Community Consultation* and which topics are currently *under discussion* (pre-proposal...

Follow the recommendations

*when you disagree, **start a debate** -do not use private rules, this only causes confusion*

Sequence Variant Nomenclature | Recommendations | Background Materials | Recent Additions | Contact Us

Current Recommendations

General	DNA	RNA
Protein	Uncertain	Checklist
Open Issues		

Background Material

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

THE HUMAN VARIOME PROJECT

HGVS HUMAN GENOME VARIATION SOCIETY

facebook & twitter

facebook

HGVS
Education

Timeline | About | Photos | Likes | Events

PEOPLE

217 likes

ABOUT

These HGVS pages will be used to discuss any subject we encounter regarding the "Recommendations for the description of sequence variants".

http://www.HGVS.org/mutnomen

PHOTOS

HGVS shared a link.
October 19

Tue. Oct. 21, 12:30-14:00, HVP Sequence Variant Description workshop ASHG, room 28A, San Diego Convention Center. What are we going to do? Discuss variant nomenclature!
After a short introduction on the basics, the
open... See More

Schedule of Events | AS
www.ashg.org
The American Society of H
Incorporated | 9650 Rock
Maryland 20814society@a
• (301) 634-7300Privacy

HGVS
October 17

Intron after stop codon
Q: how do I number a variant which is at position 13 in an intron immediately following the last nucleotide (c.876) of the stop codon?
c.*0+13C>T can not be since HGVS does not use position "0".
A: since the variant is in an intron at position 13 after nucleotide c.876 the correct description is c.876+13C>T.
Interesting to note is that in this peculiar example nucleotides in the intron are numbered like c.876+1, c.876+2, c.876+3, … c.*1-3, c.*1-2, c.*1-1.

JT den Dunnen @jtdendunnen 6d
HGVS and ISCN
HGVS made recommendations to describe variants at nucleotide level. However, first variants... fb.me/2xWBGUDly

JT den Dunnen @jtdendunnen 11 Oct
Unique indel being an inversion
Q: how to describe variant c.3821_3825delTCACTinsAGTGA, an in-frame indel... fb.me/1hJnxny03

JT den Dunnen @jtdendunnen 11 Oct
The basiscs - slide presentation .. now updated.
The slide presentation explaining the basics of the variant... fb.me/2778rhFVz

Versioning

Sequence Variant Nomenclature | Recommendations | Background Materials | Recent Additions | Contact Us | Version 15.11

Versioning

The recommendations for the description of sequence variants are designed to be ***stable***, ***meaningful***, ***memorable*** and ***unequivocal***. Still, every now and then small modifications will be required to remove inconsistencies and/or to clarify confusing conventions. In addition, the recommendations may be extended to resolve cases that were hitherto not covered. To allow users to specify up to what point they follow HGVS nomenclature, version numbers will be assigned.

Since 2015, **any change** in the recommendations receives a new **version number**. The version number will be based on the date of the change. Both in the *version list*, and on the page containing the change, the version number assigned will be clearly marked. The version number will have the format: **HGVS nomenclature *Version 15.11***, for the version accepted in 2015 ("**15**"), November ("**11**").

The current HGVS version number is shown in the top right corner of this web site ("***Version xx.xx***"). Note that the version number remains as is when only a typing error is corrected, an example added, an explanation clarified, a question answered, etc.

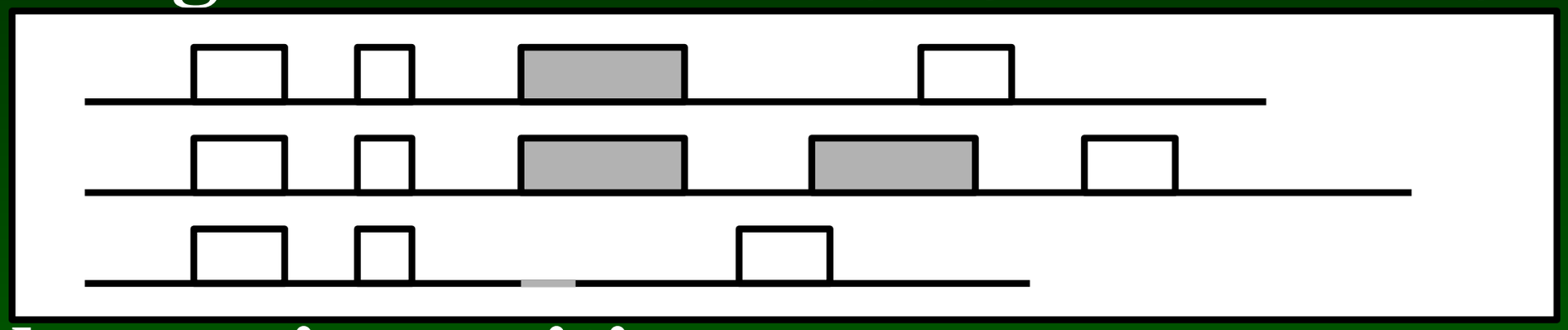
version presented is 15.11 (Nov.2015)

Variant types

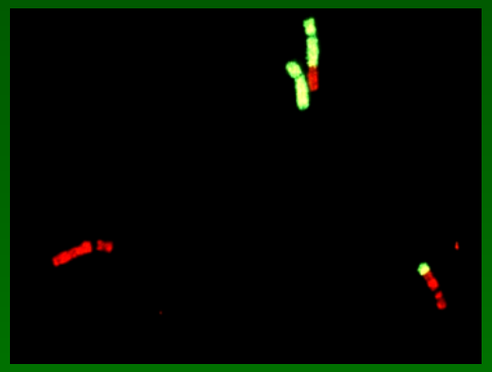
- **change in sequence**

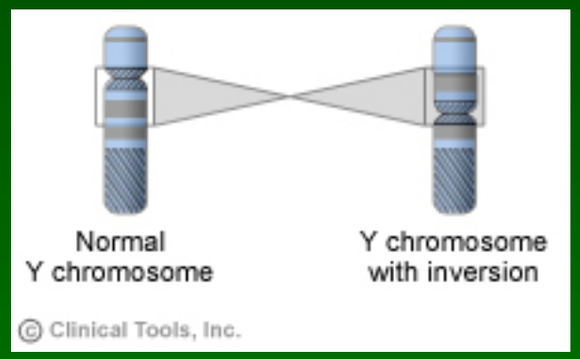
```
ACATCAGGAGAAGATGTTC  GAGACTTTGCCA
ACATCAGGAGAAGATGTTT  GAGACTTTGCCA
ACATCAGGAGAAGATGTT   GAGACTTTGCCA
ACATCAGGAGAAGATGTTCCGAGACTTTGCCA
```

- **change in amount** *(Copy Number Variation)*



- **change in position**





Structural Variation (SV)

DNA, RNA, protein

- **unique descriptions**
  prevent confusion
- **DNA**
  A, G, C, T
  g.957A>T, c.63-3T>C
- **RNA**
  a, g, c, u
  r.957a>u, r.(?), r.spl?
- **protein** *(mostly deduced)*
  three / one letter amino acid code
  ** = stop codon*
  p.(His78Gln)

Reference sequence

- use official HGNC gene symbols
- provide reference sequence
  - *covering complete sequence*
    - *largest transcript*
  - *preferably a LRG*
    - *e.g. LRG_123*
  - *give accession.version number*
    - *e.g. NM_012654.3*
- indicate type of Reference Sequence
  - *DNA*
    - *coding DNA* *c.* *genomic* *g.*
    - *mitochondrial* *m.* *non-coding RNA* *n.*
  - *RNA* *r.*
  - *protein* *p.*

The LRG

Dalgleish *et al. Genome Medicine* 2010, 2:24
http://genomemedicine.com/content/2/4/24

Genome Medicine

CORRESPONDENCE Open Access

Locus Reference Genomic sequences: an improved basis for describing human DNA variants

Raymond Dalgleish[1]*, Paul Flicek[2], Fiona Cunningham[2], Alex Astashyn[3], Raymond E Tully[3], Glenn Proctor[2], Yuan Chen[2], William M McLaren[2], Pontus Larsson[2], Brendan W Vaughan[2], Christophe Béroud[4], Glen Dobson[5], Heikki Lehväslaiho[6], Peter EM Taschner[7], Johan T den Dunnen[7], Andrew Devereau[5], Ewan Birney[2], Anthony J Brookes[1] and Donna R Maglott[3]

Abstract

As our knowledge of the complexity of gene architecture grows, and we increase our understanding of the subtleties of gene expression, the process of accurately describing disease-causing gene variants has become increasingly problematic. In part, this is due to current reference DNA sequence formats that do not fully meet present needs. Here we present the

Introduction

In 1993 Ernest Beutler
editor of the *American*
lighting the deficienci
describe DNA variants
Human Mutation invit
Tsui to produce a nome
proteins [2]. From the
years have borne witn

EDITORIAL

nature genetics

Conventional wisdom

Recent agreement on stable reference sequences for reporting human genetic variants now allows us to mandate the use of the allele naming conventions developed by the Human Genome Variation Society.

By agreement between stakeholders and two principal databases, it has been proposed (R. Dalgleish *et al.*, *Genome Med.* 2, 24, 2010, doi:10.1186/gm145) that human genetic variants be reported relative to a new set of stable reference sequences, "Locus Reference, Genomic" (LRG, pronounced "large" http://www.lrg-sequence.org/page.php). These sequences have been developed from the initial NCBI RefSeqGene concept and are provided by NCBI and EBI according to agreed rules age, resequencing and marker association studies and so keep allele descriptions commensurate with the method by which their data were generated.

The LRG reference sequences should be used in conjunction with standard HGNC gene abbreviations (http://www.genenames.org/) that we already require as a condition of publication. All human genetic variants must now be described—in abstracts and at first use—in accor-

EBI, NCBI, Gen2Phen

GEN2PHEN

Numbering residues

- **start with 1**
  - *genomic* — *1 is first nucleotide of file*
    - *no +, - or other signs*
  - *coding DNA* — *1 is A of ATG*
    - *for introns refer to genomic Reference Sequence*

- **repeated segments** — *(...CGTGTG TG A…)*
  - *assume most 3' as changed*

- **coding DNA only**
  - *5' of ATG* — *…, -3, -2, -1, A, T, G, …*
    - *no nucleotide 0*
  - *3' of stop* — **1, *2, *3, …*
    - *no nucleotide 0*
  - *intron*
    - *position between nt's 654 and 655*
    - *c.654+1, +2, +3, ……, -3, -2, c.655-1*
    - *change + to - in middle*

Numbering

- **RNA** *(deduced mostly)*

  like coding DNA

- **protein** *(deduced only)*

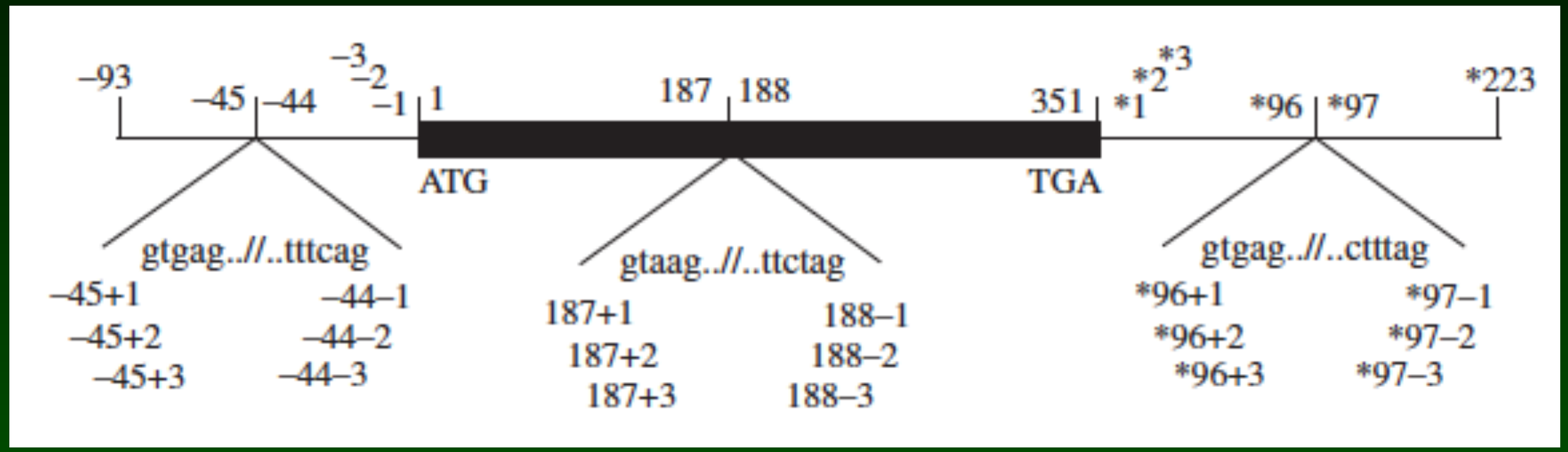
  from first to last amino acid

  rule of thumb: c. nucleotide position divided by 3 roughly gives amino acid residue

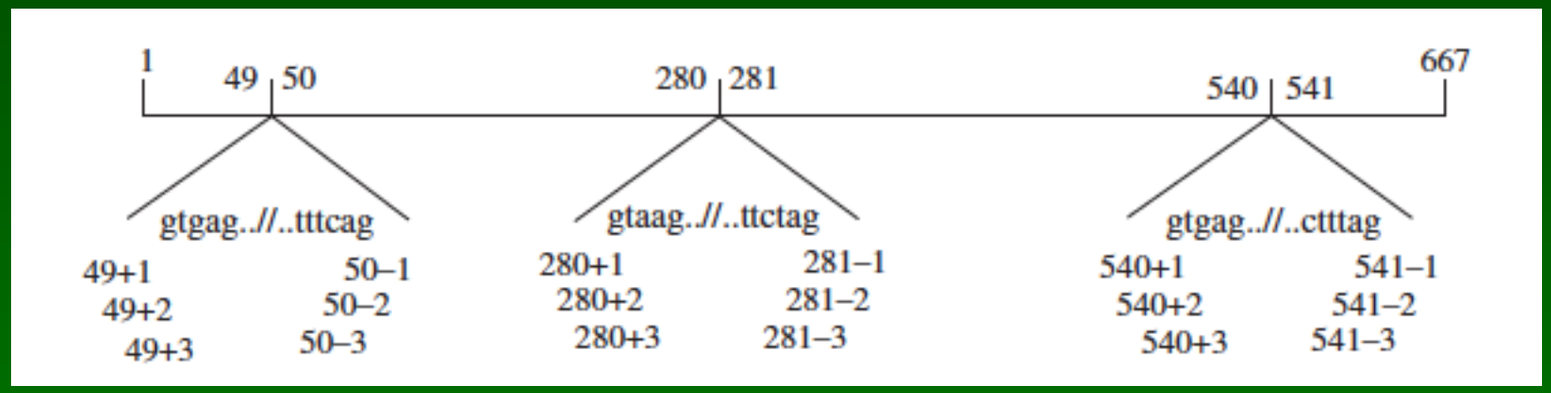
  description between parantheses

Reference Sequence

coding DNA reference sequence (c.)



non-coding DNA reference sequence (n.)



coding DNA or genomic ?

- **human genome sequence**
  - *complete*
  - *covers all transcripts*
    - *different promoters, splice variants, diff. polyA-addition, etc.*
  - *but*
    - *hg19 chr2:g.121895321_121895325del*
    - *is long & complicated*
    - *huge reference sequence files*
    - *new builds follow each other regularly*
    - *carries no understandable information*

- **coding DNA**
  - *does not cover all variants*
  - *but gives a clue towards position*

Numbering - genomic[3]

- g.12158663A>G
- g.23669859>C
- g.89112396G>A
- g.112775623C>G
- g.56569443A>T
- g.12741333T>G
- g.188153979G>C

no relation to RNA & protein

Numbering - coding DNA

- **c.1637A>G**
  protein coding region
- **c.859+12T>C**
  in intron (5' half)
- **c.2396-6G>A**
  in intron (3' half)
- **c.-23C>G**
  5' of protein coding region (5' of ATG)
- **c.*143A>T**
  3' of protein coding region (3' of stop)
- **c.-89-12T>G**
  intron in 5' UTR (5' of ATG)
- **c.-649+79G>C**
  intron in 3' UTR (3' of stop)

relation to RNA & protein

Types of variation

- **simple**

substitution	*c.123A>G*
deletion	*c.123delA*
duplication	*c.123dupA*
insertion	*c.123_124insC*
other	

conversion, inversion, translocation, transposition

- **complex**

indel	*c.123delinsGTAT*

- **combination of variants**

two alleles	*c.[123A>G];[456C>T]*
>1 per allele	*c.[123A>G;456C>T]*

Substitution

- **substitution designated by ">"**
  > not used on protein level

- **examples**

genomic	*g.54786A>T*
cDNA	*c.545A>T*
	(NM_012654.3 : c.546A>T)
RNA	*r.545a>u*
protein	*p.(Gln182Leu)*

Deletion

- **deletion**
  designated by "del"
  range indicated by "_"

- **examples**

  c.546del
  c.546delT

  c.586_591del
  c.586_591delTGGTCA, NOT c.586_591del6

  c.(780+1_781-1)_(1392+1_1393-1)del
  exon 3 to 6 deletion, breakpoint not sequenced

Duplication

- **duplication**
  designated by "dup"
  range indicated by "_"

- **examples**

  c.546dup
  c.546dupT

  c.586_591dup
  c.586_591dupTGGTCA, NOT c.586_591dup6
  do not describe as insertion

  c.(780+1_781-1)_(1392+1_1393-1)dup
  exon 3 to 6 duplication, breakpoint not sequenced
  NOT: dup should be in tandem

Insertion

- **insertion**
  designated by "ins"
  range indicated by "_"
  ! give inserted sequence

- **examples**

  c.546_547insT
  NOT c.546insT or c.547insT

  c.1086_1087insGCGTGA
  NOT c.1086_1087ins6

  c.1086_1087insAB567429.2:g.34_12567
  when large insert submit to database and
  give database accession.version number

Inversion

- **inversion**
  - *affecting at least 2 nucleotides*
  - *designated by "inv"*
  - *range indicated by "_"*

- **example**

  c.546_2031inv

  NOT c.2031_546inv

Conversion

- **conversion**
  - *affecting at least 2 nucleotides*
  - *designated by "con"*
  - *range indicated by "_"*

- **examples**

  c.546_657con917_1028

  c.546_2031conNM_023541.2:c.549_2034

Sequence repeats

- **mono-nucleotide stretches**
  g.8932A(18_23)

  c.345+28T(18_23)
  alleles 345+28T[18];[21]

() = uncertain

- **di-nucleotide stretches**
  c.1849+363CAG(13_19)
  c.1849+363_1849+365(13_19)

- **larger**
  g.532_3886(20_45)
  3.3 Kb repeat

SNVs *(SNPs)*

- **SNV's**

***at least once** give description based on genome reference sequence*

hg19 chr9:g.3901666T>C

rs12345678:T>C

dbSNP entry

Characters & codes

- **codes used**

+, -, *	
>	***substitution (nucleotide)***
_	***range***
;	***separate changes (in/between alleles)***
,	***more transcripts***
( )	***uncertain***
[ ]	***allele***
=	***equals reference sequence***
?	***unknown***
del	***deletion***
dup	***duplication***
ins	***insertion***
inv	***inversion***
con	***conversion***
ext	***extension***
fs	***frame shift***

Uncertainty breakpoints

- **Copy Number Variants**

(last-normal_first-changed) _ (last-changed_first-normal) del

BAC / PAC probe

chrX:g.(32218983_32238146)_(32984039_33252615)del
hg19

SNP-array

chrX:g.(32218983_32238146)_(32984039_33252615)del
GRCh36.p2
(rs2342234_rs3929856)_(rs10507342_rs947283)del

Uncertainty breakpoints 2

- **whole exon changes**

c.(423+1_424-1)_(631+1_632-1)del
intragenic deletion

c.(?_-79)_(631+1_632-1)del
deletion incl. 5' end

c.(423+1_424-1)_(*763_?)del
deletion incl. 3' end

c.(?_-79)_(*763_?)del
whole gene deletion, start/end undefined

describe what was actually tested

Alleles

- **allele**
  indicated by "[ ]", separated by ";"
- **2 changes, 2 alleles**
  c. [428A>G] ; [83dupG]
- **1 allele, several changes**
  c. [12C>G ; 428A>G ; 983dupG]
- **2 changes, allele unknown**
  c. 428A>G (;) 83dupG
- **special cases**
  mosaicism
  c. 428A= / A>G
  chimerism
  c. 428A= // A>G

spaces in description used for clarity only

Complex

- **deletion / insertions**

  "indel"

  c.1166_1177delinsAGT

- **descriptions may become complex**

  when only an expert understands the "code" consider database submission

  description: c.875_941delinsAC111747.1

Changes in RNA

- **description like DNA**

  r. / a, g, c, u

- **examples**

  r.283c>u
  r.0 no RNA from allele
  r.? effect unknown
  r.spl affects RNA splicing
  r.(spl?) may affect splicing
  r.283= no change
  (equals reference sequence)

  r.[=, 436_456del]
  two transcripts from 1 allele

Changes in RNA$_2$

- **one allele, 2 transcripts**
  effect on splicing not 100%

c.456+3G>C

on RNA r.[=, 436_456del]

> p.[=, Arg146_Lys152del]

Changes in protein

- **description like DNA**

p. / Ala, Cys, Gly, His, …, Ter
*p. / A, C, D, E, F, G, H, …, **

- **examples**

no RNA data
r.(?)
p.(Trp56)*

nonsense
p.Trp65** (p.W65* / p.Trp65Ter)
no stop
p.*1054Glnext*31
p.0 - no protein
p.Met1? - likely, but unknown effect
NOT p.Met1Val
fs - frame shift

Frame shifts

- **short form** *(sufficient)*

  p.Arg83fs

- **long from** *(more detail)*

  p.(Arg83Serfs*15) *(no RNA analysis)*

indicate

first amino acid changed
position
first changed amino acid
length shifted frame
(from first changed to * incl.)
do not describe del, dup, ins, etc.

do not try to include changes at DNA level

Recent additions

- **added versioning**
  - *to support users*
    - *easier to find latest changes*
    - *allows statement "following HGVS version 2.0"*
- **stricter definitions**
  - *separate different classes*
  - *added hierarchy*
    - *computer-generated description*
    - *automated error-checking (Mutalyzer)*
- **simplified use special characters**
  - *"_", ";", "+", "*", ...*
  - *improved consistency*

Acknowledgement

Presentation prepared by:

Johan den Dunnen

Human Genetics & Clinical Genetics
Leiden University Medical Center
Leiden, Nederland

chair SVD-WG

date: April 2017