

A formalized description of the standard human variant nomenclature in Extended Backus-Naur Form

Additional File 2 – Protein variant nomenclature EBNF

Jeroen F.J. Laros¹, André Blavier², Johan T. den Dunnen¹, Peter E. M. Taschner^{1§}

¹Department of Human Genetics, Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, Nederland

²Interactive Biosoftware, Rouen, France

Email: Jeroen Laros - j.f.j.laros@lumc.nl; André Blavier - ablavier@interactive-biosoftware.com; Johan den Dunnen - ddunnen@humgen.nl; Peter Taschner[§] - P.Taschner@lumc.nl

[§]Corresponding author

Extended Backus-Naur Form of protein variant nomenclature v.2.0.0^a

Basic lexemes

AA	→	AA1 AA3 'X'
AA1	→	'A' 'R' 'N' 'D' 'C' 'Q' 'E' 'G' 'H' 'I' 'L' 'K' 'M' 'F' 'P' 'S' 'T' 'W' 'Y' 'V'
AA3	→	'Ala' 'Arg' 'Asn' 'Asp' 'Cys' 'Gln' 'Glu' 'Gly' 'His' 'Ile' 'Leu' 'Lys' 'Met' 'Phe' 'Pro' 'Ser' 'Thr' 'Trp' 'Tyr' 'Val'
Number	→	[0-9] ⁺

Name → [a-zA-Z0-9_]⁺

Top-level Rule

ProteinVar → SingleVar | MultiVar

Locations

AALoc → AAPtLoc | AARange

AAPtLoc → AA PtLoc

PtLoc → ('-' | '*')? Number
| Number ('+' | '-') Number

AARange → Extent
| '(' Extent ')'

Extent → AAPtLoc '_' AAPtLoc

Reference sequences

Ref → (Name ':')? 'p.'

Single Variations

SingleVar → Ref RawVar

RawVar → Subst | Del | Dup | VarSSR
| Ins | Indel | FrameShift
| '=' | '?' | '0' | '0?'

Subst → AAPtLoc AA ('extX' '*'? Number)?
| ('Met1' | 'M1') ('?' | 'ext' '-' Number)

Del → AALoc 'del'

Dup → AALoc 'dup'

VarSSR → AALoc '(' Number '_' Number ')'

Ins → AARange 'ins' (AA⁺ | Number)

Indel → AALoc 'delins' (AA⁺ | Number)

FrameShift → ShortFS | LongFS

ShortFS → AAPtLoc 'fs'

LongFS → AAPtLoc AA 'fs' 'X' Number

Multiple Variations

MultiVar → SingleAlleleVars
| MultiAlleleVars
| UnkAlleleVars

SingleAlleleVars → Ref SingleAlleleVarSet

SingleAlleleVarSet	→	<code>‘[’ RawVar (; RawVar)⁺ (, RawVar)⁺ ‘]’</code>
MultiAlleleVars	→	<code>Ref SingleAlleleVarSet ; Ref? SingleAlleleVarSet</code>
UnkAlleleVars	→	<code>Ref ‘[’ RawVar (;) RawVar ‘]’</code>

^a Please note that although the names of the production rules in the two EBNFs may be similar, the rules themselves are different. Terminal symbols are in **bold**. See [2] for a full list of symbols and their use.